Genomic and metagenomic analyses of the skin microbiota

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Jennifer Elizabeth Kelly

September 2013

UNIVERSITY OF LIVERPOOL

Acknowledgements

I would first like to thank my parents, who have provided me with constant support and encouragement.

I am particularly grateful for the years of advice and guidance provided by my supervisors Prof. Neil Hall and Dr. Mal Horsburgh. Special thanks are reserved for our collaborators at Unilever for providing the axillary samples, in particular Sally Grimshaw, whose perseverance made the work possible, and the late David Taylor, whose unending enthusiasm regarding this project inspired and motivated me.

I would also like to thank the numerous members of the CGR and Lab H, who provided me with advice, guidance and friendship throughout my time at Liverpool, and kept me sane with welcome distractions during my writing up period.

Abstract

Following birth the skin is rapidly colonised by microorganisms that, over time, delineate into niche-specific microbial communities that often exhibit specific host-associated functions. Due to local physiological conditions, the axilla boasts a unique microbial community that has been implicated in malodour generation via the biotransformation of odourless host-secreted substrates. To more comprehensively understand the role of the axillary microbiome in malodour generation, axillary samples of subjects exhibiting high and low malodour profiles were subject to metagenomic sequencing. Metagenomics is a relatively novel whole-genome shotgun technique that utilises high-throughput sequencing to taxonomically and functionally characterise microbial communities. Prior to the axillary analysis, an *in vitro* synthetic microbial community of known composition was created and subject to metagenomic sequencing and analysis to determine which methods most accurately represent the taxonomic and functional composition of a microbial community. Additionally, to allow a more thorough understanding of the intraspecies diversity of the most abundant skin genus *Staphylococcus*, the commensal resident *Staphylococcus epidermidis* and the closely related pathogen *Staphylococcus aureus* were both subject to comparative pan-genome analysis. Utilising a direct whole-genome sequencing approach revealed that *Corynebacterium* might not dominate the axillary microbiota as predominantly as previously thought. A wide range of microbial clades were associated with high levels of axillary malodour, however only the four following species-level groups were enriched: *Corynebacterium amycolatum, Corynebacterium kroppenstedtii, Finegoldia magna* and *Kocuria rhizophila.* The characterised ability of certain corynebacterial species to generate malodorous compounds indicates that *C. amycolatum* and *C. kroppenstedtii* may play a major role towards the generation of axillary malodour. Pan-genome analysis of the most abundant skin isolate *S. epidermidis* and its relative *S. aureus* resulted in the complete description of the core genome of both species, and revealed that *S. epidermidis* exhibits a much higher degree of intra-species variability than *S. aureus*. Also, although both species occupy distinctly divergent life-styles, a large proportion of the conserved function was present in the core-genomes of both species, indicating a high degree of shared conservation. Utilisation of high-throughput sequencing technologies allowed a more in-depth analysis of the axillary microbiota and the intraspecies variability of *S. epidermidis* and *S. aureus.*

Table of Contents

List of Tables

List of Figures

CHAPTER 1

General Introduction

1.1 The skin microbiome

1.1.1 Physiological structure of the skin

Human skin is comprised of two layers: an outer stratified squamous epithelium known as the epidermis, and a sub-epidermal layer called the dermis which is tightly connected to the epidermis via a basement membrane [1]. The epidermis is composed of five layers that are each host to keratinocytes at different stages of keratinisation, the process by which keratinocytes produced in the base stratum basale layer undergo terminal differentiation to end up as corneocytes in the outer most stratum corneum (SC) layer [2]. The SC subsequently undergoes a process called desquamination, which involves the continual shedding of corneocytes and results in the complete replacement of the epidermis approximately every 48 days [3]. The skin also comprises multiple appendages and accessory structures including sweat glands, sebaceous glands and hair follicles. Eccrine glands are the most abundant type of sweat gland and continuously secrete a watery-type fluid, whilst apocrine glands are restricted to skin sites associated with hair [4,5]. Sebaceous glands are distributed ubiquitously throughout the skin and are only absent from the palms of the hands and soles of the feet [6]. They secrete a lipid-rich oily fluid described as sebum which protects the skin against water loss and maintains skin moisture [7,8].

1.1.2 Role of the skin

The surface of the skin acts as a physical barrier against invading pathogens, whilst the low pH and abundance of antimicrobial compounds such as antimicrobial peptides (AMPs), lipids and lysozyme act as a chemical barrier to inhibit initial and prolonged colonisation. The 'acid mantle' describes the mildly acidic conditions of the surface of the stratum corneum which is partly attributable to eccrine secretions [9,10]. The density of sweat and sebaceous glands, local topography and the use of cosmetic/personal hygiene products all influence the pH level of specific skin sites, however the average pH of the skin surface is thought to be approximately 5.5 [10-12]. Constitutively expressed AMPs such as dermicidin (DCD) and human β-defensin 2 (HBD-2) represent a major component of the skin's defensive strategy against pathogenic microorganisms [13]. DCD is expressed in eccrine glands and displays antimicrobial activity against certain species of the following genera: *Staphylococcus, Escherichia, Enterococcus* and

Candida [14,15]. A larger array of AMPs including HBD-1, HBD-3 and a variety of cathelicidins are expressed by keratinocytes in response to inflammatory stimuli to protect colonisation of damaged skin [16,17].

1.1.3 The skin as a microbial niche

Due to the extreme variation in physiological and topographical conditions, the skin is host to a diverse range of microbial niches which each boast distinct environmental conditions [18]. The density of sweat glands, sebaceous glands and hair follicles, along with the thickness and degree of invaginations and folds of a specific skin site define the local environmental conditions, and determine the composition of the inhabiting microflora [19]. For example, the moist, warm and sheltered conditions of the axilla are in complete contrast to the dry environment of the volar forearm which is characterised by a low level of sweat and sebaceous glands [20]. Accordingly, there is a considerable degree of variation between the microbial composition of the forearm and axillary microflora [21]. Skin sites are either categorised as dry, moist or sebaceous depending on the density of sweat and sebaceous glands, and each encourages the colonisation of different groups of bacteria [22,23]. The oxygen availability also has a major influence upon the composition of the skin microbiome, and although skin sites are predominantly aerobic, certain occluded regions such as hair follicles, the axillary vault and deeper layers of the stratum corneum can support microaerophilic or anaerobic environments [22,24]. Host and external factors such as age, gender, cosmetic use, antibiotic use, level of personal hygiene and climate also impact the taxonomic profile of the microbiota at specific skin sites [9,25].

1.1.4 The taxonomic composition of the skin microbiota

The skin is a sterile surface *in utero* that is rapidly colonised by microorganisms following birth [26,27]. The skin microbiome of the neonate is characterised by a low level of microbial diversity which increases with age and exposure, and many factors including environment, occupation, antibiotic usage and gender are thought to influence the final taxonomic composition [27]. Initial studies investigating specific components of the skin microbiome relied on culture-dependent methods, and consequently the adult skin surface was predicted to host a relatively low diversity of microbial inhabitants [28,29]. Pioneering work in the mid-1980's first utilised 16S rRNA gene sequencing for taxonomic identification and phylogenetic analysis of unknown microbial isolates [30]. This technology was rapidly applied to the characterisation of whole microbial communities by the direct cloning of PCR-amplified microbial DNA into various types of vectors, followed by chain-termination or dye-terminator sequencing [31-37]. The development of high-throughput sequencing subsequently allowed 16S rRNA characterisation

of microbial communities on much larger scale, generating the volume of data required to accurately profile entire communities [21,38,39]. As well as making large-scale 16S rRNA assays possible, the development of high-throughput sequencing also allowed the characterisation of microbial communities via a whole-genome shotgun approach, known as 'metagenomics', which involved the direct sequencing of DNA extracted from an environmental or clinical sample [40-49].

High-throughput studies have revealed the diversity of the skin microbiota is much greater than initially predicted by culture-dependent methods [21,38,50-54]. The majority of skin residents belong to the following four phyla: Actinobacteria, Firmicutes, Proteobacteria and Bacteroidetes, however the composition at species and genus levels is dependent upon the specific skin site [52,55]. Sebaceous or oily areas such as the forehead, back and face exhibit the lowest level of microbial diversity, and are usually dominated by members of the *Propionibacterium* genus, while moist sites such as the axilla, nares and navel are dominated by *Corynebacterium* and *Staphylococcus* species [21,50,55]. Dry skin sites such as the volar forearm, buttock and palm host the most diverse microbial communities and are dominated by members from all four prevalent phyla [21,50,55]. As dry sites are often exposed (forearm and palm), it has been hypothesised that the high level of diversity is a result of the frequent exposure or contact with microbially colonised material, and recent studies have demonstrated that the composition of the skin microbiota is considerably influenced by frequent and direct contact with other microbial hosts such as family members and/or pets [50,54,56].

The skin microbiome has undergone extensive characterisation via high-throughput 16S rRNA gene sequencing, however due to the low biomass level and the subsequent low yield of extracted DNA from skin isolated samples, metagenomic characterisation of this microbial community has been limited [21,23,38]. In 2013 Mathieu *et al*., presented the only application of whole-genome metagenomic sequencing of skin samples, however to circumvent the problem of low DNA yield, multiple samples from different sites and subjects were combined prior to sequencing [53]. Although this provided a comprehensive description of the major functional pathways of the skin microbiota, the study did not analyse the variation in community composition between different skin sites [53].

1.1.5 The role of the skin microbiota in health and disease

The skin microbiota has been implicated in numerous host-associated roles, many of which have a direct impact upon human health. Certain members of the skin microbiome can directly inhibit the colonisation and proliferation of pathogenic bacteria on the skin surface, either by

production of antimicrobial compounds or by stimulation of the host innate immune response [57]. The prevalent skin resident *S. epidermidis* can synthesise a range of antimicrobial molecules, such as the inhibitory serine protease Esp and the phenol-soluble modulins (PSMs) γ and δ, which inhibit the growth of the major nosocomial pathogen *S. aureus* [58,59]. Numerous studies have also revealed that the skin microflora are able to amplify the innate immune response by secreting pro-inflammatory factors in response to the presence of pathogens, which induces host expression of antimicrobial peptides [60]. *S. epidermidis* can also induce human keratinocyte expression of antimicrobial peptides HBD-2 and HBD-3 by synthesis and secretion of the recently characterised lipopeptide LP01 [61,62]. As well as responding to invading pathogens, the resident microflora can also modulate immune responses following skin injury, suppressing excess inflammation [63,64]. As well as a role in pathogen protection, actions of the skin microbiota have also been associated with the development of a number of inflammatory skin disorders. Atopic dermatitis, acne vulgaris and psoriasis are three relapsing skin disorders which have been linked to a dysbiosis of the skin microbiome [65-68]. The development of acne vulgaris has been correlated with a general increase in microbial diversity, whilst atopic dermatitis and psoriasis have been associated with elevated level of *S. aureus* and *S. epidermidis,* and *Streptococcus* colonisation, respectively [65,66].

The colonising microflora of the skin have also been implicated in the generation of body odour via the enzymatic release of malodorous compounds from odourless precursors [69,70]. Malodorous emissions are predominately emitted from the axilla and the feet, however as axillary malodour is more widespread amongst the population and often more a pronounced problem, it has been subject to more extensive research than foot odour [69,71]. Malodorous compounds responsible for axillary malodour are released from odourless precursors in sweat and sebaceous secretions via the enzymatic actions of the axillary microbiota [29,72-74]. The three groups of malodorous compounds thought to account for the majority of axillary malodour are 16-androstene steroids, short and medium chain volatile fatty acids and sulfanylalkanols [72,73,75-78]. The metabolism of odourless precursors has been attributed to fatty-acid metabolising corynebacteria, *Staphylococcus* spp. and *Propionibacterium* spp. present in the axillary microbiota, however there is a lack of *in vivo* evidence linking specific species and malodourous compounds [74,79,80].

1.1.6 *Staphylococcus*: an important skin genus

Staphylococcal species comprise the majority of the skin microbiome, and are therefore often implicated in many host-associated roles such as pathogen protection and malodour generation

[21]. They dominate moist skin sites including the nares, axilla, plantar heel (sole of the foot), umbilicus (navel) and popliteal fossa (back of knee) [81]. It was recently shown that staphylococcal colonisation begins at a very young age, and decreases in abundance with age as microbial diversity increases [27]. *S. epidermidis* and *S. aureus* are two of the most important host-associated staphylococcal species due to their respective roles in pathogen protection and disease. *S. epidermidis* is the most prevalent commensal member of the skin microbiota and has been identified as a major component of the core skin microbiome [81,82]. Accordingly, *S. epidermidis* has been implicated in a number of host-associated roles which prominently include protection against invading pathogens by secretion of antimicrobial compounds, and as a causative agent of axillary malodour via the generation of volatile fatty acids in the axilla [58,83,80]. Although a commensal member of the skin microbiota, *S. epidermidis* is also one of the leading etiological agents of indwelling-medical device associated infections, and is responsible alongside *S. aureus* for the majority of hospital-acquired infections [84]. *S. aureus* is predominantly associated with colonisation of the nasal mucosa, and is estimated to reside in the nasal cavities of ~30% of the population [85]. Skin colonisation of *S. aureus* is thought to result from an imbalance of the residential microbiota that normally inhibit its presence by the secretion of antimicrobial compounds. Its presence on the skin is often associated with the development of a variety of skin diseases including atopic dermatitis [86]. The specific inhibition of *S. aureus* growth by *S. epidermidis* secreted antimicrobial compounds highlights the antagonistic effect between the two species.

1.2 Next-generation sequencing

Most of our recent knowledge regarding microbial community structure, composition and function is a result of the incorporation of advanced molecular techniques with high-throughput sequencing platforms. The vast volume of sequence data generated by these massively-parallel sequencing machines has provided unbiased access to entire microbial communities, including previously unknown components hidden by their low abundance or uncultivable nature, and provides insights into complex intra-community and host-associated interactions. Due to their association with metagenomics, only the two main high-throughput sequencing platforms (Illumina and 454) will be reviewed in detail.

1.2.1 The history of DNA sequencing

The first automated DNA sequencer, the AB370, was released by Applied Biosystems in 1987 and utilised capillary electrophoresis and chain-termination sequencing or 'Sanger sequencing' to generate 500 kb of DNA sequence data per day [163]. This technology was rapidly developed and the current Sanger machine, the AB3730xl, utilises dye-termination sequencing to produce 2.88 mb of data per day, with an average read length of 900 bp (LIFE TECHNOLOGIES). Sanger sequencing was the primary tool used to generate the first draft of the human genome sequence, which was published in February 2001 after costing approximately US$3 billion [164]. Following the release of the human genome sequence a new phase of DNA sequencing began with the release of three high-throughput DNA sequencing technologies coined as 'next-generation' or 'second generation' sequencers. The first pyrosequencer was released in 2005 by 454, followed by the release of the Genome Analyser by Solexa in 2006 and the SOLiD (sequencing by oligo ligation detection) sequencer by Agencourt. The three technologies were bought and subsequently developed by Roche, Illumina and Applied Biosystems respectively, and along with the more recent Ion Torrent technology, comprise the main NGS platforms. Since the release of the first NGS machine the amount of data produced by NGS platforms has more than doubled every year, and the cost per base pair has more than halved every five months (http://www.genome.gov/sequencingcosts/).

1.2.2 454 pyrosequencing

Roche 454 pyrosequencing applies massively-parallel sequencing-by-synthesis to millions of amplified DNA templates immobilised onto minute capture beads [165]. The library preparation protocol begins with the nebulisation of high quality, high molecular weight genomic DNA into appropriately sized fragments and the ligation of specific adaptors and multiplexing barcodes

onto each fragment. Each DNA fragment is then ligated onto minute capture beads and amplified within a water-oil emulsion complex, known as a micro-reactor, which generates millions of clonally amplified DNA templates on each individual bead. Following enrichment each bead is deposited into a single PicoTiterPlate plate well approximately 29 μm in diameter. During the sequencing-by-synthesis reaction deoxynucleoside triphosphates (dNTPs) are sequentially flowed over the immobilised templates in the order TACG, leading to the release of pyrophosphate (PPi) if a complementary base is incorporated by DNA polymerase [166]. The presence of ATP sulfurylase and adenosine 5' phosphosulfate (APS) lead to the conversion of PPi to ATP which supplies the energy required for luciferase to oxidise luciferin [166]. This last reaction generates visible light proportional to the number of bases incorporated that is captured by a charge-coupled device (CCD) camera and output as a flowgram.

Roche released the GS FLX Titanium system in 2008, and with an upgrade to the GS-FLX+ system read lengths of up to 1,000 bp and a throughput of 700 mb can be generated in less than 24 hours. The advantages of 454 sequencing in comparison to other platforms are primarily the long read length and quick turnaround time, with automation available for many of the template preparation steps.

1.2.3 Illumina sequencing-by-synthesis

Since the release of the Genome Analyser (GA) system in 2006, Illumina has become one of the major providers of DNA sequencing technologies, and has subsequently released two additional DNA sequencing systems, the HiSeq and MiSeq (http://www.illumina.com/systems.ilmn). The HiSeq system comprises four high-throughput machines: the HiSeq 1000, HiSeq 1500, HiSeq 2000 and HiSeq 2500, which can output a maximum of six billion reads per 11 day run using a 100 bp paired-end approach. The MiSeq is a bench top sequencer aimed at smaller genome projects, allowing considerably faster run times in comparison to the higher throughput HiSeq and GAIIx machines, and accordingly has a lower maximum output of 15 million reads (http://www.illumina.com/systems.ilmn).

Following fragmentation and specific adaptor ligation of high-quality genome material, the final Illumina library is amplified via a process known as ‘cluster generation’, in which individual DNA fragments are clonally amplified by bridge-PCR [167]. DNA fragments are ligated to the solid base of a flow-cell coated in oligonucleotides complementary to the adaptors attached to the DNA fragments during library preparation. The double stranded DNA is then denatured which allows the each strand to bend over and ligate to a complementary primer on the flow-cell, forming a bridge which is subsequently extended by DNA polymerase resulting in a double-stranded bridge [168]. Subsequent denaturing and extension cycles result in the generation of clusters of clonally amplified single stranded fragments at millions of locations

within the flow-cell. During the library preparation process a distinctive barcode can be ligated to DNA fragments from each sample allowing multiplexing of multiple samples on the same flow-cell lane.

Following the bridge-PCR reaction the nucleotide sequence of each cluster is derived via a sequencing-by-synthesis reaction [169]. All four nucleotides tagged with distinct fluorescent labels are introduced to the flow-cell simultaneously and if complementary, are incorporated into the cluster fragments via a polymerase-catalysed reaction. Following an incorporation event each tile within the flow cell is imaged to detect the specific nucleotide. To ensure only one nucleotide is incorporated during any single incorporation event, the 3' OH group of all fluorescently labelled nucleotides is blocked, and following the imaging step this blocking group is chemically removed to allow subsequent incorporation and imaging cycles. The number of cycles determines the eventual read length of the fragments, which varies from 100 to 250 depending on the specific machine http://www.illumina.com/systems.ilmn).

1.3 Molecular analysis of microbial community composition

1.3.1 Microbial community ecology

In environmental niches microbial species often occur as members of highly diverse and complex microbial communities, rather than as independent free-living organisms. Extensive interspecies interactions occur within microbial communities, with many species thriving due to their syntrophic relationships with other species, in which the metabolites generated by one species are utilised by another [89]. Microbial communities often act as cohesive units, utilising the collective biochemical potential of the entire community to catalyse and regulate a myriad of processes, ranging from globally important biogeochemical systems such as nitrogen fixation to more niche specific processes such as aiding nutrient acquisition from indigestible fibres in the human gastrointestinal tract [90]. The advent of next-generation sequencing has allowed elucidation of the structure and composition of numerous microbial communities, allowing a more detailed understanding of the roles played by whole communities and individual species within many ecosystems [45-47,91,92].

1.3.2 16S rRNA gene profiling of microbial communities

The study of environmental microbial communities was first approached using culture-dependent techniques which required pure cultures to be isolated by selective media, followed by multiple physiological and biochemical tests in order to assign a taxonomic identification. This approach mediated the identification of only a small fraction of the total microbial diversity due to the large proportion of uncultivable microbial species which are estimated to account for 0.001% to 15% of the microbial community[93]. Pioneering work by Woese *et al*. and many others in the early 1980's led to the recognition that the highly conserved ribosomal RNA genes could be utilised for culture-independent typing of environmental isolates, due to the presence of variable regions which exhibited significant interspecies specificity [93-98]. The earliest studies utilising this technique involved direct extraction of 5S rRNA molecules from samples, followed by subsequent separation using electrophoresis, however the short length and separation requirement limited the effectiveness of this approach to describe more complex microbial communities [99-101].

Subsequent studies recognised the potential of the 16S rRNA gene for microbial profiling, which contains nine variable regions designated V1-V9 which are flanked by conserved regions, allowing targeted amplification of a specific region by universal primers [102]. Although

no single variable region is able to discriminate all bacterial species, utilisation of the whole gene or a carefully selected combination of regions can allow classification to genus-level and often species or strain-level. Due to its length and utility for phylogenetic resolution, the 16S rRNA gene has subsequently become the standard choice for molecular microbial ecology studies. Following PCR amplification of 16S rRNA genes from an environmental DNA sample, taxonomic profiles of communities can be reconstructed using either sequence based, fingerprint based or microarray based techniques, although the sequencing approach is most widely utilised [35,97,103-106].

1.3.2.1 *i)* Sequencing-based 16S rRNA gene profiling

The original and most extensively applied 16S rRNA gene profiling method involves the cloning of PCR amplified 16S rRNA gene fragments into plasmid vectors, which are transformed into competent cells and subsequently screened and sequenced [21,22,39,98,107-113]. Filtered and assembled 16S rRNA gene sequences are then aligned against a comprehensive 16S rRNA database such as GreenGenes or the Ribosomal Database Project (RDP) and assigned a phylogenetic classification based on a sequence similarity cut-off value [114,115]. Although clone-based 16S rRNA profiling has been extensively applied to a myriad of microbial communities, this method can only ever uncover a partial view of the complete microbial diversity of a community due to the limited number of clones that are selected and sequenced. The introduction of short read high-throughput sequencing platforms in the last few years has allowed a cloning-independent method of 16S rRNA gene profiling. This approach involves PCR amplification of one or more of the variable regions of the 16S rRNA gene using primers complementary to the flanking conserved regions, and direct sequencing of the amplified product [23,39,52,116].

This approach has been used to characterise the taxonomic profiles of a number of host-associated and environmental microbial communities [21,117,118]. An influential study by Costello *et al*. in 2009 surveyed the microbial composition of 27 human sites across four time points using 16S rRNA gene profiling, and highlighted the considerable temporal variation of the human microbiome [50]. Extensive characterisation of the human gut microbiota using this technique has revealed an association between specific microbial community compositions and certain disease states such as obesity and diabetes, and has also suggested the division of all gut microbiota into three distinct enterotypes [119-123]. Taxonomic profiling of skin-associated microbial communities revealed that individuals generally exhibit a greater degree of inter- than intra- personal variation, and that neonates are colonised immediately after birth with a low

diversity microbial community [21,27,124]. A selection of the non-human associated sites profiled using this approach include the soil, freshwater lakes, the ocean and acid mine drainage production sites [92,125-127].

1.3.2.2 *ii)* Fingerprint- and microarray- based 16S rRNA gene profiling

Fingerprint- and microarray- based techniques represent a lower cost and more rapid approach of 16S rRNA gene profiling in comparison to sequencing. Fingerprint techniques involve the direct analysis of 16S rRNA PCR products to generate a 'fingerprint' profile of a community using either terminal restriction fragment length polymorphism (t-RFLP), temperature or denaturing gradient gel electrophoresis (T/DGGE) or single-strand conformation polymorphism (SSCP), which are all based on sequence or length polymorphisms [97,128,129]. Although these techniques do not assign taxonomic identities to community members, they provide a rapid method to determine the extent of variation between multiple microbial communities, and are often combined with sequence based methods [22,97,109,110]. Microarrays comprise the attachment of often hundreds of thousands of known 16S rRNA complementary probes to a solid surface, and the subsequent hybridisation of the fluorescently labelled 16S rRNA PCR products [99]. This approach allows a large-scale reproducible analysis of microbial communities, and has been applied to numerous environmental samples [105,106]. A limitation of this technique is the inability to detect novel prokaryotic taxa, which is often a major goal of many microbial community analyses projects.

1.3.3 Whole genome analysis of microbial communities

A complementary approach to single marker-gene based analysis of microbial communities is to target the whole-genome sequences of constituent species, which in addition to providing unbiased taxonomic information, can also provide insight into the gene content of the community, allowing inferences to be made regarding the potential ecological roles of the individual species and community as a whole.

1.3.3.1 Whole genome shotgun sequencing

The techniques applied to whole-genome sequencing of microbial communities originated from whole-genome sequencing of individual isolates. Prior to the advent of next-generation sequencing platforms, whole-genome sequencing involved a shotgun-cloning approach in which genomic DNA fragments between 2-40 kb in length were cloned into fosmid or plasmid vectors and subsequently transformed into *Escherichia coli* cells. Following replication and propagation, vector DNA was extracted and sequenced using dideoxynucleotide chain-

termination sequencing. In 1995 the first bacterial whole genome sequences were generated using this approach: *Haemophilus influenzae* and *Mycoplasma genitalium* [130,131]. Although a laborious approach, this methodology resulted in the generation of over 300 bacterial genome sequences and dramatically advanced our understanding of bacterial diversity, horizontal gene transfer of mobile genetic elements and operon structure. The introduction of next-generation sequencing platforms allowed generation of shotgun sequencing libraries without the requirement for prior cloning. The general procedure for preparing a library from extracted DNA involves enzymatic or mechanical fragmentation of high quality genomic DNA, followed by end-repair, platform-specific adaptor ligation, size selection and PCR amplification [488]. The ligation of sample-specific barcodes also allows multiple genomes to be analysed in the same sequencing run [488]. Although the read length of next-generation sequencing platforms is significantly shorter than Sanger generated reads, the considerable increase in data throughput allows a deeper coverage of each base, increasing the overall accuracy of the genome sequence. Sequencing reads are subsequently filtered for low quality reads and PCR artefacts and assembled into longer contiguous sequences using either a *de novo* approach or by aligning reads against a reference sequence. The most prevalent tools used for single genome assembly are Newbler, Velvet, SOAPdenovo, Euler-SR, IDBA and Phrap [132-135], http://www.phrap.org/phredphrapconsed.html). Assembled contigs are then subject to annotation, which involves the identification of putative coding regions and subsequent functional prediction, which is often achieved via a BLAST homology search against a non-redundant public database or databases of hidden-markov models (HMMs), which are built from multiple protein sequence alignments [136,137].

The large number of bacterial whole-genome sequences generated in recent years has highlighted the considerable degree of sequence variation that exists between strains of the same species [138-140]. Although marker-gene based profiling methods are often able to characterise the phylogenetic identification of an environmental isolate to species-level, the often substantial variation in gene content is not taken into consideration. Although two isolates may share very similar 16S rRNA gene sequences, their gene content and therefore functional capability may differ dramatically. Therefore it has become common practice to classify a species based on its 'pan-genome' which is defined as the sum of all unique genes associated with all strains, which are further categorised into 'core' and 'dispensable' genes which describe genes present in every isolate and genes not present in every isolate [141].

1.3.3.2 Metagenomics

Metagenomics is the application of whole genome sequencing to microbial communities utilising DNA extracted directly from an environmental sample. In an attempt to isolate novel biosynthetic pathways from uncultured residents of the soil microbiota in 1998, Handelsman *et al*. first applied shotgun sequencing to mixed microbial DNA extracted directly from an environmental sample, and termed the analysis 'metagenomics' [142]. Earlier metagenomic studies utilised either bacterial artificial chromosomes (BACs) or fosmids as vectors for large segments of extracted environmental DNA which were subsequently transformed into *Escherichia coli* cells and sequenced using dideoxynucleotide chain-termination sequencing [142,146-148]. The subsequent introduction of high-throughput sequencing platforms allowed direct sequencing of fragmented environmental DNA and resulted in the generation of short-read datasets containing hundreds of thousands, to millions of reads [39,41,48,126,149-151].

This novel method of microbial community analysis made it possible to not only understand the taxonomic composition of a microbial community, but also to understand its functional potential by identifying associated genes, biochemical pathways and metabolic functions. Utilising a metagenomic approach also provides a route to the uncultivable portion of the microbial community, and many studies have shown it is possible to re-construct whole-genome sequences from low-complexity microbial communities [143]. Tyson *et al*. reconstructed the genomes of *Leptospirillum* group II and *Ferroplasma* type II from an acid mine drainage (AMD) biofilm whilst Strous *et al*. successfully assembled the whole genome sequence of *Kuenenia stuttgartiensis* from shotgun data of a complex bioreactor community [143,144]. Metagenomics studies have also isolated millions of genes from microbial communities with no known homologs, exhibiting the vast degree of uncharacterised functional diversity present within certain microbial communities [126]. Finally, metagenomic analysis allows the association of specific taxa, genes or microbial compositions with certain niches (e.g. soil, mouth, skin, gut) or environmental conditions (e.g. periodontal disease, axillary malodour, psoriasis), allowing the identification of the possible etiological factors. For example in 2006 Turnbaugh *et al*. observed specific differences between the taxonomic composition of the gut microbiota in lean and obese individuals, suggesting gut bacteria influence the metabolic potential of the host, and in 2011 enriched genes for antimicrobial peptides were identified in oral communities isolated from subjects with no prior history of dental caries, suggesting a probiotic effect by the commensal flora [46,145].

Metagenomic sampling of any environment invariably results in the co-extraction of contaminating organic or non-organic material, and isolation of high quality DNA without inhibitory contaminants is often the most technically challenging step in any metagenomics project. Different environmental samples require various pre-treatments to remove the niche-

specific contaminants. Marine samples are often subject to a filtration step prior to nucleic acid extraction to concentrate samples and remove contaminating material, however careful selection of filter size can also allow enrichment for either viral, bacterial or eukaryotic cells [126,152]. The extraction of DNA from soil samples utilises one of two approaches: either by direct cell lysis of the raw soil sample or by initial isolation of prokaryotic cells from the soil sample using a density gradient and subsequent cell lysis [153,154]. Although the former approach generates a higher yield of un-biased DNA it often results in the co-extraction of polyphenols and humic acid that can interfere with downstream enzymatic steps and therefore requires extensive purification which may result in loss of nucleic acid material [125,155,156]. Contamination of host-associated genomic material is a major issue for human isolated samples such as stool, skin washes, clinical biopsies and oral samples, and often results in datasets comprised predominantly of human originating reads, which need to be removed prior to data analysis [46]. Removal of human cells or genomic material prior to sequencing often results in the associated loss of prokaryotic cells leading to a biased representation of the microbial community, therefore samples are not often filtered for contamination, leading to the removal of a substantial proportion of the subsequent sequencing data [157,158]. Recently a number of commercial kits have emerged from New England Biolabs (NEB), Molzym and MoBio which claim to allow depletion of human DNA from mixed human/microbial genomic samples, however a direct comparison between treated and untreated samples is yet to be done (https://www.neb.com/products/e2612-nebnext-microbiome-dna-enrichment-kit, http://www.molzym.com/service/licensing/molysis.html, http://www.mobio.com/secondary_dna_clean-up/powerclean-dna-clean-up-kit.html)

Another major technical challenge faced when preparing human-associated metagenomic libraries is the extremely low yield of DNA obtained from most host-associated samples. In general between 500 ng and 1 μg of high-quality genomic DNA is required to generate a next-generation sequencing library, while the typical yield isolated from a host-associated sample can be as little as 1 pg. Many studies have utilised a random whole-genome amplification method known as multiple displacement amplification (MDA) to increase overall DNA yield, however there are known biases associated with this approach including preferential amplification of low GC sequence [159-162]. Illumina have recently developed a novel method of library preparation which utilises *in vitro* transposition to generate libraries from as little as 1ng of starting material, although these kits have yet to be validated using bacterial community samples (http://www.illumina.com/products/nextera_dna_sample_prep_kit.ilmn).

1.4 Computational analysis of metagenomic data

The analysis of high-throughout metagenomic datasets presents a considerable computational challenge due to the complexity of the data and the extremely large number of reads generated by the current sequencing technologies. The recent popularity of metagenomics had led to the development of a myriad of new tools that can cope with large data volumes and can decipher the convoluted datasets to answer the fundamental questions regarding microbial community structure and function.

There are numerous steps required to transform raw data into an accurate prediction of the taxonomic and functional content of a microbial community, and for each step a myriad of different computational tools are available. The main steps involved in a typical metagenomic analysis project include but are not limited to: *i)* filtering raw-datasets for low-quality reads, sequencing artefacts and environmental contamination, *ii)* taxonomically classifying filtered reads to estimate the phylogenetic composition of the sample, *iii)* assembly of filtered reads into longer contiguous sequences to allow more accurate gene finding and whole-genome recreation and *iiii*) identification of predicted genes to estimate the functional potential of the community.

1.4.1 Pre-processing of metagenomic data

Filtering raw metagenomic data prior to downstream analysis is imperative to generate accurate conclusions regarding the taxonomic and functional content of a microbial community [170]. Errors can originate from a variety of sources including the sequencing process; which can introduce poor quality reads, error containing reads and artificially duplicated reads, and the sampling process; which can result in environmental contamination including host-associated material [171-173].

Sequencing errors result from either the calling of an incorrect base, the incorporation of one or more additional bases (insertion(s)) or the omission of one or more bases (deletion(s)) during the base-calling process. Additionally there are platform-specific error characteristics such as an increased error rate in homopolymer regions in 454 reads and a substantial reduction in read quality towards the 3' end of Illumina generated reads. Artificially duplicated reads are an additional sequencing artifact, although they originate from the library preparation step of the sequencing process [42]. As failure to remove artificially duplicated and erroneous reads may lead

to mis-estimations of taxa and gene abundance, there are a variety of freely available tools that can filter metagenomic datasets. The most popular tools include PRINSEQ, NGS QC TOOLKIT, FastQC and FASTX-Toolkit (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/, http://hannonlab.cshl.edu/fastx_toolkit/) [301,427].

Sample-associated contamination such as soil particles, faecal material and human tissue is often filtered prior to DNA extraction, however certain contaminants, including human cells, cannot be filtered out without imposing considerable biases upon the microbial cells within the sample. Therefore there is often a considerable proportion of sample-associated genetic material within the final dataset that needs to be removed prior to downstream analyses. DeconSeq and BMTagger are two alignment-based tools which allow filtering of human contaminating material from metagenomic datasets [291,489].

1.4.2 Metagenomic assembly

Due to the short read length of sequencing platforms usually employed for metagenomic studies, it is beneficial to assemble metagenomic data into longer contiguous sequences prior to further analysis. *De novo* assembly of Sanger-based genomic data was based on the well established Overlap Layout Consensus (OLC) method, which performs an all verses all read comparison to identify overlapping reads and subsequently assemble a consensus sequence [174-176]. Due to the computational cost associated with performing the all verses all step, it is not feasible to apply this method to next-generation data. Therefore, in the past five years a novel approach to genome assembly has been developed based on the *de Bruijn* graph approach, which is able to handle the extensive volumes of data generated by high-throughput platforms [177].

The *de Bruijn* graph approach involves breaking down reads into unique substrings of the defined length *k,* and finding overlaps between these k-mers to build contiguous sequences [134]. Within a *de Bruijn* graph k-mers represent the edges whilst each node represents a subsequence of length k-1 that is common to the two connected edges. Contiguous sequences are assembled from the *de Bruijn* graph by following the simplest non-branching path which satisfies the requirement that every edge is visited once. The length of the k-mer is a crucial parameter to ensure an accurate assembly, if too short then multiple k-mers will overlap with each other and introduce branches into the graph, which results in the production of very short contigs. Selecting longer k-mer lengths will usually improve an assembly, however if the length is too

long then overlaps may not be found between k-mers particularly if the reads have a high error rate resulting in gaps in the assembly. Larger k-mers lengths can also distinguish smaller repeats, which often introduce branches into graphs. Therefore to ensure accurate assemblies, k-mer lengths are often less than the total read length but longer than 50% of the read length. Due to the explosion of next-generation sequencing a myriad of *de novo* assembly tools have been developed which employ the *de Bruijn* graph based approach, however the most prominently used are Velvet, SOAPdenovo, AbySS, Euler and ALLPATHS [132,133,178-181].

Application of a *de Bruijn* graph based assembly algorithm to metagenomic reads is complicated by the specific characteristics of metagenomic data in comparison to single genome data. The presence of species at varying abundances within the microbial community results in uneven coverage distributions, which violates the clonality assumption required by most assembly tools. Also the co-existence of closely related species with homogenous regions within their genomes, and strains with almost identical genomes results in sequences that cannot be differentiated between species and strains. This introduces branches into the *de Bruijn* graph and causes the generation of short and/or inaccurate consensus sequences. Recently, two *de Bruijn* graph based tools: IDBA-UD and MetaVelvet, were developed to explicitly deal with the uneven coverage distribution characteristic of metagenomic data by identifying sub-graphs within the central *de Bruijn* graph based on a k-mer coverage histogram that represent closely related or single genomes, and resolving them [182,183]. Another tool often utilised for metagenomic assembly is SOAP*denovo,* a single genome *de novo* assembly tool which produces adequate assemblies from metagenomic data [159,181,184,185].

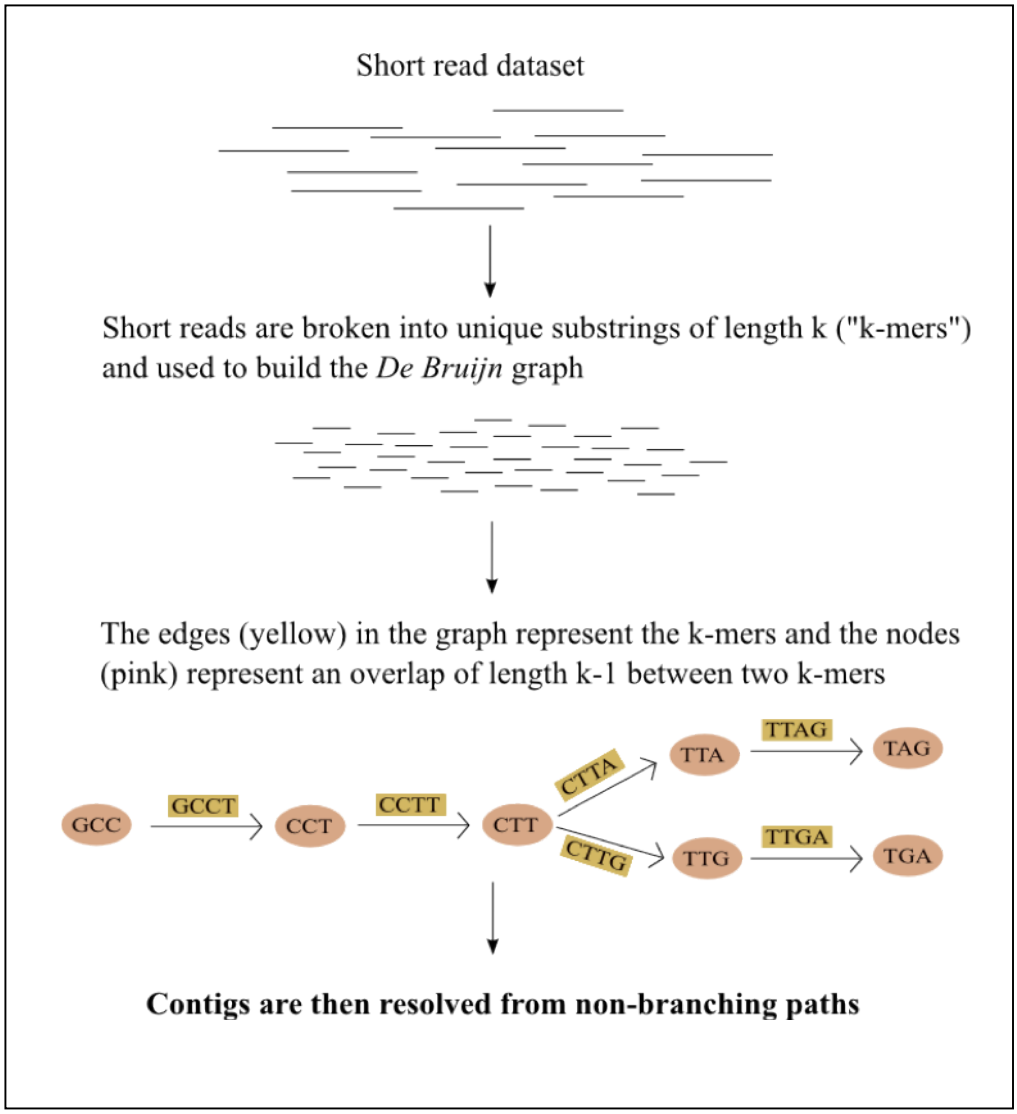


Figure 1.1. Overview of the *De Bruijn* method of whole-genome assembly. Reads are initially broken into unique sub-sequences of length k (k=4 is used in this figure for simplicity) which are subsequently used to build a *De Bruijn* graph comprising edges and nodes. Each edge is represented by a k-mer whilst a node is introduced if two k-mers share an overlap of length k-1. As represented in this figure branches or forks are introduced into the graph if there are multiple overlaps. Contigs are generated by following the simplest path that contains no branches/forks.

1.4.3 Taxonomic reconstruction of microbial communities ('Binning')

Accurately predicting the identity and relative abundance of microbial clades within a microbial community is a fundamental aim of any metagenomics project, and is often referred to as 'binning'. The majority of microbial community taxonomic profiling has been carried out using a 16S rRNA approach which involves high-throughput sequencing of either one or more hypervariable regions, clustering of sequences into operational taxonomic units (OTUs) and taxonomic classification against a database of 16S rRNA gene sequences [186]. Although this approach has significantly broadened our understanding of the microbial composition and diversity of a wide range of environments, there are certain limitations associated with the 16S rRNA amplicon method which may result in an incomplete or biased estimate of the taxonomic profile of a community, including copy number variation between species, PCR artefacts and the low resolution power at species level [187].

Taxonomic classification methods for whole genomic data are usually categorised as either composition-based or comparison-based, however there are a subset of tools that also combine both approaches [188,189]. Composition-based tools such as Phymm and Phylopythia use sequence characteristics such as GC content and k-mer variation to distinguish reads originating from different taxonomic clades, however the short read lengths of current next-generation platforms limits the effectiveness of these methods with metagenomic data [188,190]. Comparison or alignment based methods involve homology searches against a reference database and are either based on BLAST/BLAT or hidden markov model (HMM) searches [136,137,191].

BLAST (Basic Local Alignment Search Tool) and BLAT (BLAST-like Alignment Tool) are two similar alignment algorithms that allow the identification of local regions of similarity between nucleotide or amino acid sequences [136,191]. Both utilise a user defined seed length to generate initial short matches that are subsequently extended to high scoring segment pairs (HSPs), however they differ in their speed, computational requirement and ability to detect distantly related sequences. BLAT is quicker at generating alignments as it indexes the reference database and scans the query sequence, rather than indexing the query sequence and scanning the reference database as BLAST does [191]. In addition, BLAST calculates all possible mismatches in the query sequence and indexes them, requiring a considerable amount of extra time and computational power [136].

Hidden markov models (HMM) are statistical models representing multiple sequence alignments of protein families, domains or motifs[137]. Each HMM represents the sequence

conservation within a protein domain or family by calculating the probability of observing specific amino acids at different positions. A major advantage of utilising a HMM-based alignment in comparison to a pair wise BLAST or BLAT search is the ability to detect very distantly related proteins by identifying highly conserved residues within the multiple sequence alignment [137].

BLAST/BLAT homology searches are the basis of many tools: MG-RAST employs a BLAT search against the M5 non-redundant (M5NR) database and assigns taxonomic classifications based on the best-hit method, MEGAN extracts the lowest common ancestor (LCA) of the top scoring hits against the NCBI *nr* database and SO-rt ITEMS performs an additional reciprocal BLAST step to refine the taxonomic classification [192-194]. CARMA and WebCARMA are HMM-search based tools which identify conserved protein families and domains within the metagenomic query sequences to generate phylogenetic classifications [195,196]. CARMA3 is an extension of CARMA which utilises both BLAST and HMM based homology searches to carry out taxonomic classification of metagenomic data [194].

As an alternative to aligning metagenomic datasets against exhaustive databases, a subset of tools exploits marker genes with high discriminatory power to predict the taxonomic profile of microbial communities. As the databases are much smaller, this approach often requires substantially less computational power, allows much faster taxonomic classifications and has been shown to outperform some of the more traditional classification methods [197,198]. MetaPhlAn and MetaPhyler are currently the only two tools to apply this method to taxonomic profiling of metagenomic datasets. The MetaPhlAn marker gene database contains a set of genes specific to every microbial clade with available reference sequences, and therefore can make very high resolution taxonomic classifications, whilst MetaPhyler relies on a set of 31 phylogenetic marker genes that have been extracted from all sequenced genomes [197,198].

1.4.4 Functional annotation of metagenomic datasets

In addition to understanding the taxonomic composition of a microbial community, the major motivation behind metagenomic sequencing is the ability to identify genes and pathways in order to generate hypotheses regarding the roles played by communities or species within specific environments, often host-associated, and novel microbial functions.

Reconstructing the functional profile of a microbial community begins with identifying putative genomic features within the assembled or unassembled dataset, which encompasses protein

coding genes (CDS), CRISPR repeats and non-coding RNAs. Most programs use sequence characteristics such as codon usage to distinguish coding/non-coding regions, and a number of tools have been designed which specifically deal with the erroneous and short read nature of metagenomic datasets including FragGeneScan, MetaGeneMark, MetaGeneAnnotator, Metagene and Orphelia [199-203]. Identified features are then compared against one or more reference databases, which are either based on protein signatures, orthologous protein families or pathways and subsystems, and are queried by either a variant of the BLAST algorithm such as RPS-BLAST, BLASTP, BLASTX, the faster alternative BLAT or the hidden markov model based algorithm HMMER [136,191,204]. As no single database represent all functional information, the annotation process usually involves the integration of annotations from multiple databases.

Protein-signature based databases include Pfam, PROSITE, PRINTS, CATH-Gene3D and TIGRFAMs, and represent multiple sequence alignments of either protein families, domains or functional sites represented by a variety of protein signatures including hidden markov models (HMM), patterns, profiles and fingerprints [205-209] . Protein signature-based functional classification allows the identification of very distantly related homologs and is the most appropriate method for annotation of unassembled reads, due to the representation of shorter protein signatures such as functional domains and sequence features [210].

The clusters of orthologous groups (COGs) and non-supervised orthologous groups (NOGs) databases comprise groups of orthologous proteins derived from at least three separate lineages [211]. As orthologs evolve from a common ancestral sequence they often perform the same or similar biochemical function, allowing functional inference of any query sequences generating significant alignments [212].

The pathway and subsystems databases SEED and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) represent collections of subsystems and pathway maps that are utilised to predict the molecular interaction networks present within a microbial communities [213,214]. The KEGG Orthology (KO) database contains all proteins involved in the associated pathways grouped into orthologous clusters [215]. Assigning homologous query sequences with KO terms allows generation of the corresponding pathway maps.

In an effort to standardise and simplify the complex process of annotating metagenomic datasets, a number of web-based and locally installable pipelines have been developed which allow functional annotation and metabolic reconstruction by combining multiple tools and databases. The most prominent pipelines include CAMERA, CoMet, the EBI Metagenomic Archive, IMG/M, MEGAN4, METAREP, MG-RAST, RAMMCAP, Smash Community and

WebMGA [192,193,216-222]. Most of the above tools utilise a wide range of databases to generate a comprehensive annotation of the submitted dataset. IMG/M and MG-RAST are two of the most commonly used web-based metagenomic analysis servers and at the time of writing contained over 2,000 and 13,000 publically available metagenomic datasets respectively. Both utilise all of the previously described databases for annotation, and additionally include homology searches of all publically available metagenomic datasets. Certain tools are more focused upon a specific annotation approach, for example the EBI Metagenomic Archive uses the protein-signature based tool InterProScan v5.0, which searches a subset of the InterPro protein-signature databases to generate functional predictions, whilst MEGAN4 focuses upon the prediction of molecular networks within microbial communities by employing the SEED and KEGG classification databases.

1.5 Aims and objectives of the thesis

Due to the low biomass of skin-associated microbial samples, no cutaneous microbial community has been subject to characterisation utilising a whole-genome shotgun approach. Therefore, our current knowledge regarding the skin microbiota is based on taxonomic profiling studies only, and understanding of the functional potential of most skin microbial communities is limited. The axillary microbiota is a dense skin-associated microbial community implicated in the generation of axillary malodour via the biotransformation of secreted host precursors into malodourous compounds. The aim of **chapter three** is to characterise the axillary microbiome using a metagenomics approach, by generating Illumina sequencing libraries from ultra-low yields of microbial community DNA. In addition to generating an unbiased taxonomic profile of this community, utilising a whole-genome approach will allow the gene content of the community to be identified, allowing predictions to be made regarding the potential biochemical pathways involved in malodour generation. If successful, this will represent the first whole-genome metagenomic characterisation of a skin-associated microbial community, and will allow subsequent profiling of other low-yield microbial communities using similar techniques.

Due to the myriad of computational tools available for the analysis of metagenomic data, it is essential to understand which tool to select in order to generate the most precise representation of the microbial community from raw sequencing reads. Therefore, to allow the most accurate taxonomic and functional profiling of the axillary microbiome, **chapter two** aims to benchmark a number of popular metagenomics analysis tools by generating an *in vitro* synthetic microbial community of known composition. Validation using a synthetic microbial community, in which the exact taxonomic composition is known, instead of a genuine microbial community, will

ensure that any differences between tools are a direct result of the analysis method and not of the uncharacterised nature of the community.

Within microbial communities exist many strains of the same species, which can vary considerably in gene content and therefore may exhibit contrasting functional characteristics. To predict the functional potential of skin-associated communities, it is necessary to generate a more comprehensive understanding of the gene variation that exists between skin isolates of the same species. Therefore, **chapter four** aims to characterise the total intraspecies diversity of the two skin-associated species *S. epidermidis* and *S. aureus*, by constructing their pan-genomes using all available strains. Due to their contrasting lifestyles, an additional aim is to identify species-specific gene content via a direct comparison of species-wide conserved genes.

The ultimate aim of this thesis is to generate a more comprehensive understanding of the skin microbiome by utilising the recent advances made in high-throughput sequencing technologies, allowing a more in-depth characterisation of microbial inhabitants than ever before.

CHAPTER 2

Assessing the accuracy of numerous metagenomic analysis tools by utilising an *in vitro* simulated microbial community sequenced from a range of DNA yields

2.1 Introduction

2.1.1 Metagenomic analysis

Whole-genome shotgun metagenomics is a powerful method of characterising the taxonomic composition and functional capabilities of entire microbial communities, without the need for any prior cultivation. The declining costs of generating high-throughput sequencing datasets and the release of a myriad of standalone and web-based metagenomic analysis tools has led to metagenomics experiencing an explosion in popularity in the past ten years, and as of September 2013, MG-RAST hosted over 90,000 metagenomes and 32 trillion base pairs of data ([192], http://metagenomics.anl.gov/). This approach has been applied to a wide range of environmental habitats including the ocean, human gastro-intestinal tract, acid-mine drainage ponds, soil and the oral cavity, allowing a more comprehensive understanding of microbial community ecology than ever before [45,126,150,156]. Correspondingly, there has been a dramatic increase in the number of freely available informatics tools developed for the various stages of a metagenomic analysis [223].

2.1.2 Metagenomic assembly tools: MetaVelvet and IDBA-UD

Due to the short read length of current next-generation sequencing platforms, the assembly of metagenomic data into longer contiguous sequences considerably improves the ability to identify longer genomic features such as protein-coding genes and RNA genes in downstream analyses. However, specific characteristics of metagenomic data including non-uniform sequencing depth and the presence of closely related strains and species void the requirements of most single-genome assembly algorithms and therefore assembly generally results in the production of inaccurate and short contigs [132,178]. Currently MetaVelvet and IDBA-UD are the only two assembly tools specifically developed for the *de novo* assembly of metagenomic data exhibiting uneven coverage depths [182,183]. Both approaches utilise a *de Bruijn* graph method which involves splitting reads into unique sequences of length *k*, and building a graph from these k-mers comprised of nodes and vertices, in which vertices (or edges) represent the k-mer sequence and nodes represent overlaps of length k-1 shared between multiple vertices [134]. Based

on differing coverage depths the main *de Bruijn* graph is then separated into distinct sub-graphs that represent single taxonomic groups. Contiguous sequences are then produced from the simplest paths in the local graphs. In an attempt to reduce the number of gaps, IDBA-UD utilises an iterative k-mer strategy, beginning the assembly with a short k-mer length and using the assembled contigs as reads for subsequent iterations with longer k-mer lengths [183]. Both tools have been compared against a range of alternative assembly programs such as Velvet, Velvet-SC, SOAP*denovo* and Meta-IDBA, and have emerged as the most accurate for metagenomic assembly, however a direct comparison between the two has never been carried out.

2.1.3 Inferring taxonomic composition from metagenomic data

As generating an accurate prediction of species diversity and abundance from metagenomic data is central to any metagenomics study, there are an abundance of taxonomic analysis tools utilising numerous techniques. Most composition-based phylogenetic classification methods developed for use with whole-genomic shotgun data utilise one of two homology-based approaches, either by aligning the reads against extensive protein and RNA reference databases and inferring taxonomy from the phylogenetic classification of either the best hit or lowest common ancestor of the top hits, or by utilising a reduced database comprising a carefully selected collection of microbial clade-specific marker genes. Tools employing the former approach include MG-RAST, MEGAN, CARMA, WebCARMA, CAMERA, IMG/M, whilst the marker gene approach is a more recent technique and is therefore utilised by fewer tools, namely MetaPhlAn and MetaPhyler [192-194,196-198,216,218]. An advantage of the marker-gene approach is a significant reduction in the computational power required due to the reduced size of the dataset, however as only a small proportion of the genome is utilised for alignment, it is possible certain microbial clades may be missed if those sequences are not adequately represented within the dataset.

MG-RAST is one of the most popular web-based metagenomic analysis tools, and has been used to discern the taxonomic and functional composition of microbial communities isolated from a myriad of environments, including the swine gut, human mouth and soil [45,125,150]. MG-RAST uses BLAT to perform an extensive homology search of all predicted protein coding and RNA genes against a myriad of databases including the NCBI-nr, SEED and RDP, and extracts hits exhibiting a significant level of similarity to assign a taxonomic classification to each read [115,191,213,224,225]. It subsequently performs normalisation of the raw counts using total counts, to predict the relative abundance of each microbial clade within the dataset, and allows creation

and export of tables, pie charts, PCoA plots, boxplots and heatmaps to represent the phylogenetic composition of the microbial community [192]. MetaPhlAn utilises an alternative approach to MG-RAST, aligning query sequences against a small database of clade-specific marker genes to generate taxonomic profiles. The selected marker genes are highly discriminatory, allowing the unambiguous classification of reads to a high-taxonomic rank, often species-level [198]. Although utilising marker-gene databases to generate taxonomic profiles of microbial communities has not been extensively applied to metagenomic datasets, this approach has the potential to dramatically improve the accuracy of taxonomic prediction in comparison to utilising alternative profiling methods due to the clade-specificity and low volume of the marker genes [85].

2.1.4 Utilising artificial datasets to compare metagenomic analysis tools

As no fully annotated metagenome is available as a reference, a popular approach utilised for comparative analysis of multiple metagenomic analysis tools is the creation of an artificial metagenomic dataset. Artificial metagenomes are generated following one of two approaches: *i*) by utilising an *in silico* approach, which involves the combination of either simulated or genuine sequencing reads, or *ii*) by combining either cellular or genomic material to create an *in vitro* synthetic microbial community followed by DNA extraction (if required) and high-throughput sequencing [226-230]. Although *in silico* simulations provide valuable information they cannot encompass sequencing-dependent errors or characteristics that may affect the composition of a genuine metagenomic dataset, and it is generally agreed that the *in vitro* approach generates a more accurate representation of a genuine metagenomic dataset. Simulated microbial communities have also been utilised to compare high-throughput sequencing platforms and to benchmark a variety of metagenomic analysis tools [226,227,231,232].

2.1.5 Approaches to accessing low yield metagenomic samples

The extraction of insufficient yields of microbial DNA from environmental samples is a major factor limiting the metagenomic characterisation of many microbial communities. A popular approach to circumvent the problem of a low DNA yield is to increase the amount of genomic DNA prior to sequencing using a whole genome amplification method (WGA). Earlier WGA methods such as degenerate oligonucleotide PCR (DOP-PCR) and primer-extension PCR (PEP) were shown to be heavily effected by secondary DNA structure leading to uneven amplification of loci across the genome and short fragment lengths and making them highly unsuitable for metagenomic sequencing [233,234]. A more recent WGA technique not affected by secondary DNA

structures is multiple-displacement amplification (MDA), which involves rolling-circle amplification of genomic DNA via random hexamer primers and the proofreading DNA polymerase from the bacteriophage Phi29 [235]. It is claimed by manufacturers that products such as GenomiPhi (GE Healthcare) and Repli-G (Qiagen) which perform MDA, carry out unbiased amplification of whole genome DNA, amplifying the template evenly, and subsequently this approach has been used to amplify extracted bacterial and viral DNA from a range of habitats including coral, human respiratory tract, skin and infant stool samples [43,49,159,162]. Although MDA efficiently increases the genomic DNA yield from the ng range to the μg range making it possible to generate sequencing libraries, it has also been associated with significant amplification bias, partly due to the formation of chimeric fragments [236]. Subsequently numerous studies have demonstrated that the use of MDA skews taxonomic abundance and often leads to preferential amplification of certain species [237,238].

An alternative approach to sequencing low-yield samples is the use of transposition-based library preparation kits such as Nextera or Nextera XT [239,240] (ILLUMINA). Standard library preparation for most next-generation sequencing platforms involves the same fundamental steps with slight alterations depending on the specific machine and library preparation kit, and usually involves approximately one day of hands-on lab time. Initially, high-quality genomic DNA is fragmented into appropriately sized fragments by sonication, nebulisation or shearing. The fragmented DNA then undergoes end-repair followed by ligation of platform specific adaptors. These steps are interceded by many rounds of purification and size selection of the fragmented DNA and the process is finalised by quality checking and quantification of the final library. Due to the numerous steps involved there is considerable loss of template DNA from start to finish, with the majority being lost during the fragmentation stage. Nextera-based technologies allow a library to be constructed from a lower DNA yield by negating the requirement for a mechanical fragmentation step and by reducing the total number of steps involved from starting genomic DNA to final library. Nextera and Nextera XT kits allow a complete sequencing-ready library to be constructed from 50 ng and 1 ng of starting genomic DNA respectively [239].

2.1.6 Nextera and Nextera XT technologies

The Nextera and Nextera XT library preparation kits are based on *in vitro* transposition using a mutated hyperactive Tn5 transposase (Tnp) encoded by the Tn5 transposon [241]. Two inverted 19 bp repeats (IS50 elements) flank the transposon and are required for function [242]. When free 19 bp ‘transposon ends’ or ‘transposition recognition motifs’ (TRMs) are used in conjunction with

the Tn5 transposase it results in fragmentation of the target DNA and ligation of the transposon ends to the 5' end of the target DNA. Nextera and Nextera XT kits take advantage of this function plus the use of a hyperactive mutated form of the Tn5 transposase to fragment genomic DNA and ligate platform-specific adaptors using limited cycle PCR (Illumina). The Nextera/XT protocol begins with the addition of a 'transposome' complex consisting of free transposon ends and the Tn5 transposase to the un-fragmented genomic DNA. The transposome complex makes double stranded breaks in the genomic DNA and ligates the free transposon end oligonucleotide to the 5' end of the fragmented DNA. Platform specific primers are then ligated to both ends of the fragmented DNA using limited-cycle PCR.

2.1.7 Development of the Nextera method

To allow sequencing from picogram levels of DNA an additional library preparation protocol was developed by Parkinson *et al*. that is a slight modification of the Nextera transposition-mediated fragmentation method [240]. The authors demonstrated that using picogram amounts of genomic DNA to prepare an E. coli K-12 Illumina library produced sequence data comparable in terms of coverage to data generated by a standard 1 μg Illumina library[240]. The modifications applied to this method include the addition of oligonucleotides complementary to the 19 bp transposon ends during the tagmentation step, which results in fragmented DNA with oligonucleotides annealed to the 19 bp transposon sequence at both ends of the DNA fragment. The oligonucleotides also have integrated type IIG restriction endonuclease recognition sites, and the subsequent addition of restriction endonucleases cleave the transposon end and the 10 bp primer leaving a 2 bp 3' overhang at both ends of the fragments. Due the short length of the restriction sites they are also likely to be present in the template DNA fragments leading to loss of a percentage of fragments when the enzymes are added. To overcome this the ultra-low input method uses four different restriction endonucleases in four separate reactions and pools the samples. PCR is then used to anneal platform specific adaptors to the fragmented DNA and to enrich the library.

Nextera and Nextera XT library preparation kits have never been applied to microbial community samples. Marine *et al*. utilised the Nextera kit to prepare both Illumina and 454 libraries from a viral community of known taxonomic abundance and diversity and found that the viral abundance rankings differed from expected with the six lowest abundance viruses ranked in the expected order and the three top abundant viruses ranked incorrectly [160]. The transposition-based library preparation kits have also been applied to a limited number of pure

culture microbial samples and human samples which concluded that the use of Nextera kits does not adversely effect coverage apart from in GC or AT rich regions [240,241] .

2.1.8 Aims of the chapter

Understanding which metagenomic analysis tools most accurately represent the functional and taxonomic composition of a microbial community is essential to generate the correct conclusions regarding the potential role of the community within a specific environment. Therefore the main aim of this chapter was to assess the accuracy of a selection of metagenomic analysis tools responsible for taxonomic profiling, assembly and functional annotation by creating an *in vitro* simulated microbial community comprising eleven bacterial species. As subsequent work in this thesis involves metagenomic sequencing of numerous axillary microbial communities, the results of this chapter will benefit the subsequent analysis by allowing selection of the most appropriate metagenomic analysis tools.

Also, as applying transposition-based library preparation methods to metagenomic sequencing would open the door to the sequencing and analysis of microbial communities exhibiting low DNA yields, an additional aim of this chapter was to assess whether these techniques could be applied to microbial community samples by utilising Nextera, Nextera-XT and Parkinson kits to generate libraries from synthetic microbial community samples comprising very low amounts of DNA.

2.2 Methods

2.2.1 Creating the synthetic microbial community

2.2.1.1 Cell culture

Eleven bacterial species were selected to form the synthetic microbial community and comprised a combination of high and low GC species with genome sizes varying from 1.8 mb to 2.6 mb (Table 2.1). Species were selected based on their predicted relative abundance in the axilla in an attempt simulate a genuine axillary microbial community (unpublished data, Unilever). All strains were sourced from one of the following culture collections: the National Collection of Type Cultures (NCTC), Collection of Institut Pasteur (CIP), the Health Protection Agency (HPA) or the German Collection of Microorganisms and Cell Cultures (DSMZ). Each organism was supplied as lyophilised cells and re-suspended in the appropriate media according to culture collection instructions. Following re-suspension cultures were incubated at the appropriate atmospheric and temperature conditions according to the organism-specific protocol. After incubation the optical density was measured at a wavelength of 600nm and cell suspensions measuring an OD_{600nm} of 1 were prepared using sterile TE buffer, aliquoted into 1 ml volumes and stored at -80°C until needed.

Table 2.1. Strain and genome information for each species used to compose the synthetic community. *For those species with no sequenced reference, genome size and GC content are estimated from the genome size of the closest relative with a publically available genome sequence.

Strain ID	Culture Collection ID	Genome Size (mb)	GC Content (%)
Staphylococcus epidermidis ATCC 12228	NCTC11128	2.56	32.1
Peptinophilus harei ACS-146-V-Sch2b	DSM10020	1.84	34.44
Propionibacterium granulosum VPI 0507, DSM 20700	DSM 20700	~2.5*	~60.1*
Corynebacterium amycolatum SK46	DSM0105	2.51	58.6
Corynebacterium mucifaciens CIP 105129	CIP105129	~2.5-2.6*	~53*
Corynebacterium tuberculostearicum SK141	CIP107291	2.4	60
Staphylococcus lugdunensis HKU09-01	CIP103642	2.7	33.9
Staphylococcus hominis SK119	NCTC9810	2.23	31.3
Anaerococcus octavius NCTC 9810	NCTC9810	~1.7-2.0*	~33-36*
Corynebacterium appendicis CIP 107643	CIP107643	~2.5-2.6*	~53*
Finegoldia magna ACS-171-V-Col3	DSM20470	1.83	32

2.2.1.2 DNA extraction

Frozen samples were thawed on ice and centrifuged to allow removal of as much supernatant as possible. Pelleted cells were then re-suspended in 200-500 µl TE buffer pH 8.0, depending on the volume of supernatant remaining, followed by the addition 4 µl of lysozyme at 250 U μl^{-1} (Ready-Lyse, EPICENTRE). Samples were incubated in Pathogen Lysis L tubes and incubated at 37°C for 18 h with shaking at a rate of 300 rpm (QIAGEN). After incubation bead beating was carried out on a FastPrep cell disruptor for 40 s at a speed of 6 m/s (meters/second), then repeated with samples held on ice for 5 minutes in-between (ZYMO RESEARCH). DNA was extracted with the automated QIAsymphony DNA extraction robot using the Qiagen Virus/Pathogen DNA extraction kit and the ComplexFix 800 protocol with carrier RNA substituted for sterile molecular grade water (QIAGEN). DNA was eluted in 60 µl sterile molecular grade water and the three samples per organism were pooled. DNA was quantified with the Qubit dsDNA high sensitivity (HS) kit with the Qubit 2.0 spectophotometer (INVITROGEN), and purity was checked with the ThermoScientific Nanodrop 2000

(THERMOSCENTIFIC). DNA was run on a 2% agarose gel prior to sequencing to check for degradation and RNA contamination.

2.2.1.3 Constructing the synthetic microbial community

The construction of the synthetic microbial community sample and the preparation of all sequencing libraries was carried out by Dr. Linda D'amore (CGR, The University of Liverpool).

The synthetic microbial community sample was prepared by combining the same number of DNA molecules from each organism. The number of DNA molecules per μl was calculated by first multiplying the concentration in ng μl^{-1} by Avogadro's number (6.022×10^{22}), which is the number of molecules per mole, then dividing that number by the average molecular weight of 2 nucleotides (656.6×10^{9}), and multiplying by the genome size. To create the standard mock community sample 1.5×10^{9} molecules of DNA from each organism was combined. The standard sample was then diluted to appropriate concentrations to create samples at a range of decreasing yields. Quantification was carried out using the Qubit Quant-IT dsDNA high-sensitivity (HS) assay kit (INVITROGEN).

Illumina- and 454- specific standard libraries were constructed from 500 ng of microbial community DNA each, following manufacturers instructions (ILLUMINA, ROCHE) (Fig 2.1). Following manufacturers recommendations 50 ng of community DNA was then used to create a standard Nextera library for each platform (ILLUMINA) (Fig 2.1). To understand if the Nextera tagmentation process was compatible with a lower yield of input DNA, two lower-yield Nextera libraries were also constructed for each platform using 500 pg and 50 pg of DNA following manufactures instructions (ILLUMINA) (Fig 2.1). For the Illumina platform only an altered version of the Nextera protocol developed by Parkinson was used to create libraries from 500 pg and 50 pg of genomic material [240]. Finally, again only for the Illumina platform, three replicate Nextera XT libraries made were made from 1 ng of DNA following manufacturers instructions (ILLUMINA) (Fig 2.1).

Following library preparation 1 μl of each final library was checked using an Agilent BioAnalyser on a high sensitivity chip (HS) to quantify the library and check for correct DNA fragmentation (AGILENT). Additionally 1 μl of the final library was quantified using the Qubit dsDNA high sensitivity (HS) kit and measured on the Qubit 2.0 spectophotometer (INVITROGEN). All libraries were transferred to the Centre for Genomic Research (CGR) at the University of Liverpool for subsequent cluster generation and high-throughput sequencing

on the Illumina GAIIx, or emulsion-PCR amplification and sequencing on the 454 FLX+ machine.

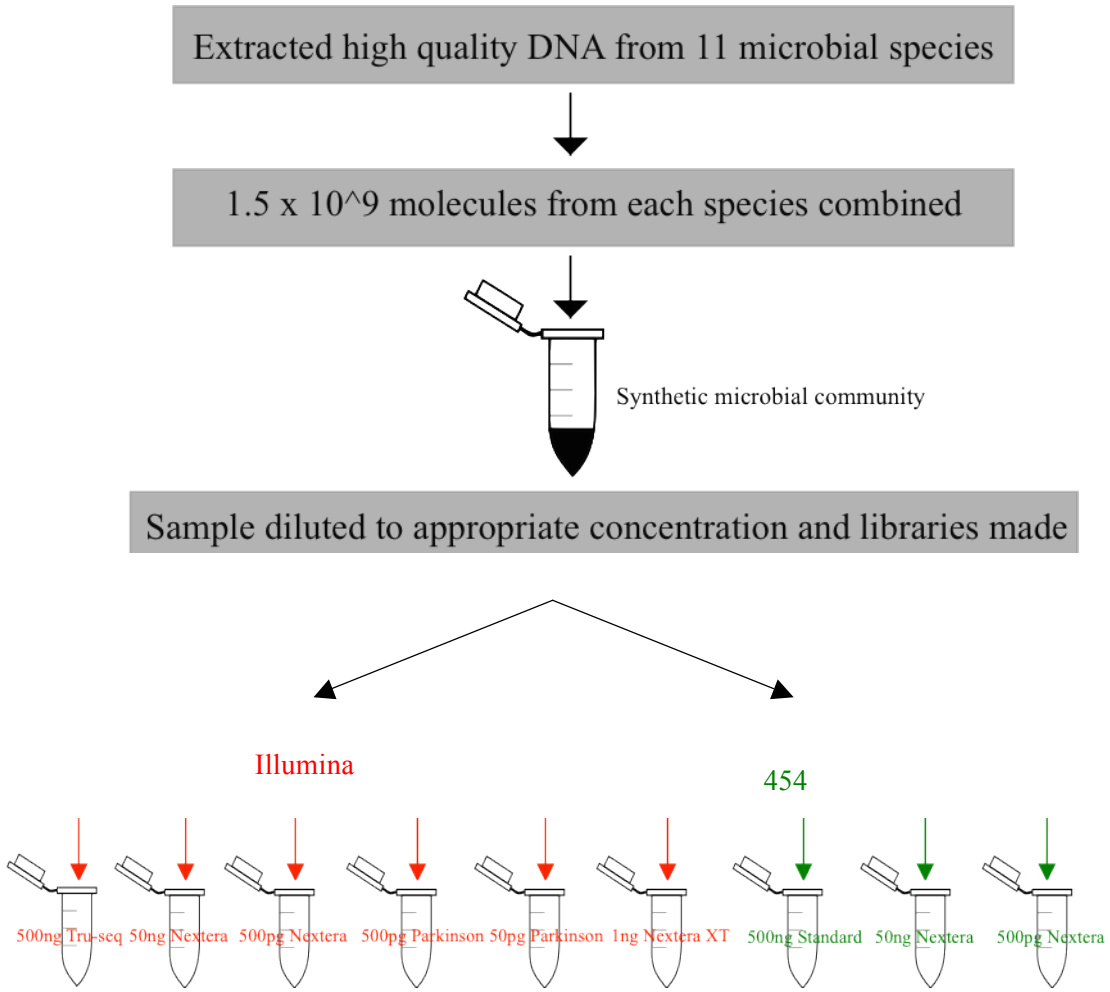


Figure 2.1. Protocol utilised to generate all synthetic microbial samples. A platform-specific standard 500 ng library, 50 ng Nextera library and 500 pg. Nextera library were constructed for all platforms, whilst 1 ng Nextera XT libraries were constructed for the Illumina platform only.

2.2.2 Generating the reference genome sequences

To allow subsequent comparative analyses of the shotgun datasets, the genome sequences of all eleven reference genomes were required. As seven out of eleven of the microbial community members had published genome sequences only the following four members required whole genome sequencing: *A. octavius, P. granulosum, C. mucifaciens* and *C. appendicis*.

2.2.2.1 Whole genome sequencing

P. granulosum and *C. mucifaciens* were subject to DNA sequencing using the 454 Titanium chemistry whilst *A. octavius* and *C. appendicis* were sequenced using the newer 454 FLX+ chemistry. Titanium and FLX+ compatible libraries were generated using 500 ng of purified genomic DNA using the standard manufactures protocol (ROCHE).

Following library preparation 1 µl of the final library was checked using an Agilent BioAnalyser on a high sensitivity chip (HS) to quantify the library and check for correct DNA fragmentation (AGILENT). Additionally 1 µl of the final library was quantified using the Qubit dsDNA high sensitivity (HS) kit and measured on the Qubit 2.0 spectophotometer (INVITROGEN). All libraries were transferred to the Centre for Genomic Research (CGR) at the University of Liverpool for subsequent emulsion-PCR amplification and sequencing on the 454 FLX+ or Titanium machine.

2.2.2.2 Assembly and annotation of reference genomes

Raw flowgram files were filtered for adaptors using the SFF tools package and fasta and quality files were obtained for each sample (ROCHE). Data was filtered for short and low-quality bases using the 454 quality filtering script from the NGS QC Toolkit software package using a minimum phred quality score of 20 over at least 70% of the length of the read (454QC.pl) [427]. Filtered reads were then assembled with the Roche *De Novo* Assembler version 2.6 using the default settings which were as follows: minimum overlap length, 40, minimum overlap identity, 90, alignment identity score, 2 and seed length, 16 (http://www.454.com/analysis-software/). To allow subsequent comparisons to the annotated shotgun datasets, assembled contigs were annotated using both the Integrated Microbial Genomes (IMG) web server, and the protein-signature-based tool InterProScan v5.0 (https://img.jgi.doe.gov/cgi-bin/er/main.cgi) [257].

The IMG-ER gene finding protocol includes a BLAST search against the IMG non-redundant rRNA database (consisting of 16S rRNA, 23S rRNA and 5S rRNA genes) to identify rRNA genes, then tRNAScan-SE-1.23, CRISPR recognition tool (CRT) and the Sanger Institute script rfam_scan are used to search for tRNAs, CRISPR sequences and other RNAs respectively [244-246]. The final step of gene finding involves identification of protein coding genes using the program GeneMark [247]. Functional annotation begins with an RPS-BLAST search against the conserved domains (CDD)[247] and PRIAM databases followed by a filtered hidden markov model search against the Pfam and TIGRfam databases and finally a BLASTp search against the IMG database [204,248-250].

InterProScan v5.0 was also utilised to generate functional annotations of all genomes. Prior to annotation, coding regions (CDS) were predicted by FragGeneScan v.1.15, functional classifications were then assigned to predicted CDS by InterProScan v5.0 which utilised the following databases: Pfam, TIGRfam, PRINTS, PROSITE patterns and Gene3d from the InterPro release 31.0 [199,205-209,257].

2.2.3 Bioinformatic analysis

2.2.3.1 Filtering raw datasets

During the Nextera library preparation protocol a 19 bp inverted repeat (transposon end) is annealed to either end of each DNA fragment. Due to the position of the platform-specific sequencing primers in relation to this transposon end, when a Nextera library is sequenced on a 454 machine the sequencing reaction begins prior to the transposon sequence, resulting in the final read containing ~38 bp of transposon sequence that must be removed before further analysis. When the Nextera library is sequenced on an Illumina machine, due to the sequencing adaptors being positioned internally to the transposon sequence, the sequencing begins after the transposon and the reads do not require trimming. Therefore, the 19 bp transposon sequences were removed from 454 datasets using a bespoke script that utilises BLAST (*blasttrim*, Appendix B Table 1).

Paired-end Illumina generated reads were filtered for low-quality bases using an Illumina-specific quality filtering script from the NGS QC Toolkit software package using a minimum phred quality score of 20 over at least 70% of the read (*IlluQC.pl*) [427]. The software filters paired-end data simultaneously, producing forward and reverse filtered fastq files and a singleton fastq file containing reads whose pair did not pass the quality filter. Transposon

trimmed 454 datasets were filtered using a 454 quality filtering script from the NGS QC Toolkit package using the same parameters (*454QC.pl*) [427]. Additionally, bases with a phred quality score below 20 were trimmed from the 3' end of all Illumina filtered reads using the NGS QC Toolkit software (*TrimmingReads.pl*) [427].

The most precise way to identify artificially duplicated reads is to use paired-end information, which was generated for all Illumina datasets, as both forward and reverse reads should be identical if the sequence is an artificial duplicate. As there is no software able to remove duplicates using paired-end information, a bespoke perl script was used to append forward and reverse reads, and forward and reverse quality strings into single reads and single quality strings (*mergeRead_fastq.pl*, Appendix B Table 1). The merged reads were then filtered for exact duplicates and reverse complement exact duplicates using the de-replication function of PRINSEQ [301].The resulting de-replicated fastq file was split back into separate forward read and reverse read fastq files using an additional bespoke perl script (*splitRead_fastq.pl*, Appendix B Table 1). Single-end reads generated by 454 platforms were also filtered for artificial duplicates using the de-replication option of the standalone version of PRINSEQ [301]. Exact duplicates, reverse compliment exact duplicates, 5' duplicates, 3' duplicates, and reverse compliment 5'/3' duplicates were removed.

2.2.3.2 Mapping datasets to reference genomes

All filtered Illumina datasets were mapped to the set of reference genomes using BWA v0.5.9-r16 [251]. The reference genome database was first indexed using the IS algorithm of the BWA index module. The filtered Illumina datasets were then aligned to the reference genomes using the sampe algorithm of BWA using the default parameters (maximum insert size 500). Uniquely mapping reads were extracted from the output BAM file to remove reads that aligned to multiple locations and genomes. The number of reads uniquely mapping to each of the 11 reference genomes was calculated using the idxstats module of SAMtools v0.1.17 [252]. Coverage statistics were calculated from the uniquely mapping reads using a bespoke perl script (*coverageStatsSplitByChr_v2.pl*, Appendix B Table 1). 454 datasets were mapped using the Roche GS Reference Mapper v2.6, using default parameters (http://www.454.com/products/analysis-software/). Bespoke perl scripts were subsequently used to extract mapping statistics (*extract_full_partial.pl* and *extract_coverage.pl*, Appendix B Table 1). Uniformity of coverage across each genome was examined by extracting coverage per base using the mpileup module of SAMtools v0.1.17 [252].

To understand if the abundance distributions of any of the datasets follow the expected model, a chi-squared goodness of fit analysis was carried out between the expected and observed distributions for all datasets.

2.2.3.3 Metagenomic Assembly

All Illumina datasets were assembled using MetaVelvet and IDBA-UD, two *de bruijn* graph based *de-novo* metagenomic assemblers [182,183]. MetaVelvet requires the user to declare the k-mer size, and as this parameter can have a drastic effect on the outcome of the assembly optimal k-mer size was determined prior to assembly using a selection of the datasets (data not shown). As the k-mer length is recommended to be no shorter than half the read length the k-mer range tested was 55-63 in incremental steps of two, and the optimal k-mer size was found to be 55. MetaVelvet also requires the insert size and its standard deviation of each dataset, which were both calculated from the BAM file generated during mapping using the CollectInsertSizeMetrics.jar module of Picard Tools v1.79 (http://picard.sourceforge.net). All filtered datasets were then assembled with IDBA-UD using the optional 'pre-correction' which speeds up the assembly by correcting erroneous reads in high-depth regions [183].

All quality-filtered 454 generated datasets were assembled using the Roche GS *De Novo* Assembler v2.8 using the default minimum overlap length of 40bp and minimum alignment identity of 90% (http://www.454.com/products/analysis-software/).

2.2.3.4 Evaluating contig accuracy

Chimeric contigs arise from the co-assembly of reads originating from different genomes into a single contig and are a recognised problem of metagenomic assembly due to the presence of closely related strains and species in microbial communities. In this project the definition of a chimeric contig is one that contains uniquely mapping reads originating from more than one reference genome [227]. To determine whether contigs were chimeric or non-chimeric, reads were mapped to both the assembled contigs and the reference genomes using BWA as previously described, and a reference ID and contig ID was obtained for each uniquely mapping read. This information was then used by a a bespoke perl script to determine which contigs contained reads originating from more than one reference, and therefore could be classified as chimeric, and which contigs contained reads originating from only one reference, and therefore could be classified as non-chimeric (*chimericity_of_assembly.sh*, Appendix B Table 1). For all contigs identified as chimeric the number of reads not originating from the reference that the majority

of the reads in that contig originate from was calculated, and this was converted into a percentage of the total reads in that contig and known as the 'degree of chimericity' [226,227].

2.2.3.5 Functional annotation of metagenomic datasets

Functional annotation was carried out using the two web-hosted annotation services: IMG/M and the EBI Metagenomic Service (https://img.jgi.doe.gov/, https://www.ebi.ac.uk/metagenomics/) [218]. The EBI Metagenomic Service accepts raw unassembled reads or assembled contigs from any platform whilst IMG/M-ER accepts raw unassembled reads from only the 454 platform and requires data generated by higher-throughput platforms such as Illumina to be submitted as assembled contigs.

All assembled Illumina and 454 contigs were submitted to the Integrated Microbial Genomes Metagenomic Expert Review (IMG/M) website for functional annotation. The IMG/M gene finding protocol began with the identification of CRISPR elements using a combination of two programs, the CRISPR recognition tool (CRT) and PILER-CR [245,253]. Non-coding RNAs were then found using tRNAScan-SE-1.2 to identify tRNAs and internally developed rRNA models to identify rRNAs [218,244]. Protein-coding genes were discovered using a conjunction of four gene finding programs: GeneMark, Metagene, Prodigal and FragGeneScan [199,202,247,254]. Functional annotation began with a search against the Pfam-A database followed by an RPS-BLAST search against the database of position-specific scoring matrices (PSSM) for COGs with an e-value cut-off of 0.01 [205,211]. A similarity search against the non-redundant IMG database was then carried out to generate assignment of KEGG Orthology terms (KO) and enzyme commission (EC) numbers [214].

In collaboration with the InterPro development team the raw 500 ng, 50 ng Nextera and 500 pg Nextera datasets from each platform were submitted to the EBI Metagenomic Service for annotation. As the EBI Metagenomic Service employs its own quality-filtering pipeline, the un-filtered datasets were used. Low-quality ends and remaining adaptor sequences were initially trimmed using the BioPython SeqIO package, followed by removal of low-quality reads and those less than 100bp in length [255]. Duplicate and repeat sequences were filtered using UCLUST v. 1.1.579 and RepeatMasker v3.2.2 respectively (http://www.repeatmasker.org) [256]. Coding regions (CDS) were then predicted by FragGeneScan v.1.15 and functional annotation of predicted CDS was achieved with InterProScan 5.0 using a selection of databases including Pfam, TIGRfam, PRINTS, PROSITE patterns and Gene3d from the InterPro release 31.0 [199,205-209,257].

2.2.3.6 Estimating functional annotation accuracy

To determine which dataset most accurately represented the functional profile of the microbial community the abundance of each COG group or InterPro entry was calculated for each dataset and correlated against the expected COG/InterPro entry abundances from annotation of the reference genomes. Pearson's correlation coefficients and p-values were calculated for each comparison using the statistical package R (R Development Core Team, 2008, http://www.R-project.org). A custom Perl script was also used to identify false-positive and false-negative COG groups/InterPro entries, and to determine which COG groups/InterPro entries were over-represented, under-represented or correctly represented with respect to the reference annotations.

2.2.3.7 Taxonomic classification

To understand how accurately the phylogenetic composition of the microbial community was predicted using existing computational tools, taxonomic profiles of all datasets were generated using two tools, MetaPhlAn and MG-RAST [192,198].

MetaPhlAn uses an alignment-based approach to map unassembled reads against a pre-computed set of clade-specific marker genes using BowTie2 [198]. The set of clade-specific marker genes covers all sequenced archaeal and bacterial phylogenies and were selected for their collective ability to distinguish microbial clades at a high taxonomic resolution. A clade-specific marker gene is defined as a gene ubiquitously present in one clade and absent from all others, therefore allowing unambiguous classification of reads to specific clades. As the number of marker genes per clade only represents a fraction of the whole genome, only a small proportion of the reads will map to the database. Relative abundance of each clade is calculated by normalising the total length of all read hits by the length of the marker genes, therefore reporting the relative abundance as a function of the total number of cells rather than the total number of reads [198]. When mapping fastq-formatted input files a non-local hit policy was invoked to avoid overtly sensitive hits that may result in false positive classifications (--bt2_ps very-sensitive).

Raw, unassembled datasets were uploaded to the MG-RAST website using the default filtering parameters and analysed using the MG-RAST v3.2 pipeline [192]. The MG-RAST server employs a specific quality control pipeline which involved the removal of reads comprising more than 5 bases with a phred quality score below 15, followed by the 3' trimming of bases with a phred quality score below 15. The quality-filtered datasets were then de-replicated which removed all but one read from a collection of reads whose first 50 bp were identical. Following quality-

control MG-RAST then identified putative protein coding genes using FragGeneScan, followed by functional annotation via a BLAT or BLAST similarity search against the M5 non-redundant protein database (M5NR), an amalgamation of the following sequence database: Gene Ontology, Greengenes, KEGG, NCBI-nr, RDP, SEED, SILVA, UniProt and eggNOG [114,115,213,214,224,258-261]. Sequence hits were then assigned a taxonomic classification using the lowest common ancestor approach with a maximum e-value of 1e-5, a minimum percentage identity of 75 and a minimum alignment length of 15. Taxonomic abundance was estimated by calculating the number of annotated features with predicted protein or RNA genes annotated to species and genus levels for all datasets. Relative abundance of each taxonomic classification was calculated by normalising abundance values against the total number of annotated features with respect to each individual dataset. Classifications with relative abundances over 0.1% and 0.5% were then extracted for further analyses.

2.3 Results and discussion

2.3.1 Creating the reference genome dataset

To allow comparative analysis of the metagenomic datasets, whole genomes sequences of all eleven species were required. The genome sequences of seven species were already publically available, therefore only *Propionibacterium granulosum* DSM 20700, *Corynebacterium mucifaciens* CIP 105129, *Anaerococcus octavius* DSM 11663 and *Corynebacterium appendicis* CIP 10764 were subject to whole-genome sequencing. *De novo* assembly of filtered reads generated high-quality assemblies for all datasets, with 99% of bases in each assembly at a quality level of Q40 or higher, which represents a miscall once every 10,000 bases.

A commonly used statistic used to measure the length of assembled contigs is the N50 length, which is defined as 'the length at which if contigs of that length or over were examined, half of the bases in the entire assembly would be represented'. Long contigs were produced for *C. mucifaciens* and *A. octavius* datasets, which exhibited N50 lengths over 150 kbp and 180 kbp respectively, whilst *P. granulosum* and *C. appendicis* generated considerably shorter assemblies with N50 lengths over 14 kbp (Table 2.2). The variation in contig length is likely to be a result of the considerable differences observed in read number between datasets, as an increased read coverage allows more accurate resolution of overlaps and a reduction of gaps during the Overlap Layout Consensus (OLC) assembly. Although multiplexed samples were pooled in eqimolar quantities prior to sequencing is it not uncommon to experience up to a 10-fold difference in data volume as seen in these datasets. Despite the varied coverage depth observed between assemblies, contigs represented at least 80% of each reference genome (Table 2.2).

Table 2.2. Assembly statistics for the four sequenced genomes assembled using the Roche GS *De Novo* Assembler tool v2.6. *Percentage is in relation to the total number of reads. **Percentage of genome covered by contigs is based on an estimated genome size calculated from the genome length of the closest relative with a publically available genome sequence.

	Total no. Bases (mb)	**Assembled Reads (%)***	**No. Contigs**	**N50 Length**	**Percentage Coverage (%)****
P. granulosum	19	97	245	14,765	86
C. mucifaciens	58.1	97	43	182,396	84-87
A. octavius	149.2	98	47	156,060	95-100
C. appendicis	19.8	97	265	16,052	86-89

2.3.2 Sequencing the synthetic microbial community

To simulate the problem of sequencing microbial communities from very low DNA yields, fourteen synthetic microbial communities were created with DNA yields ranging from 500 ng to 0.05 ng (Table 2.3). Libraries were prepared using a variety of methodologies selected for their ability to generate low input libraries, and sequenced on two high-throughput sequencing platforms (Table 2.3). Henceforth each dataset will be referred to by the dataset IDs listed in Table 2.3, which represent the sequencing platform, method of library preparation and DNA yield (Table 2.3).

Table 2.3. Platform, library and DNA yield information for each of the sequenced shotgun datasets.

Dataset ID	Platform	Library Protocol	Quantity of Input DNA (ng)
Illu500	Illumina	Tru-Seq	500
IlluN50	Illumina	Nextera	50
IlluN0.5	Illumina	Nextera	0.5
IlluP0.5	Illumina	Parkinson	0.5
IlluP0.05	Illumina	Parkinson	0.05
IlluX1	Illumina	Nextera XT	1
IlluX2	Illumina	Nextera XT	1
IlluX3	Illumina	Nextera XT	1
Pyro500	454	Rapid Library	500
PyroN50	454	Nextera	50
PyroN0.5	454	Nextera	0.5

2.3.3 Removal of low-quality reads and artificial duplicates

As the presence of low quality reads and artificial duplicates can impose considerable biases upon downstream analyses, all fourteen datasets were subject to a rigorous quality-filtering pipeline. The number of filtered reads was normalised with respect to the total number of reads to allow inter-sample comparisons. When all datasets were examined, between 6.7% and 75.2% of reads were classified as either low-quality and/or artificial duplicates and removed (Fig 2.2). As extremely low-yield DNA samples are often associated with a reduced complexity, it is possible that generating libraries from as little as 50 pg of DNA may have introduced a higher proportion of artificial duplicates into the resulting datasets. However, for both platforms there was no significant correlation between the amount of DNA used to generate the library and the proportion of reads classified as low-quality/artificial duplicates, indicating that a substantially lower level of input DNA does not negatively impact upon the read quality (Pearson's product-moment correlation test: $p > 0.05$).

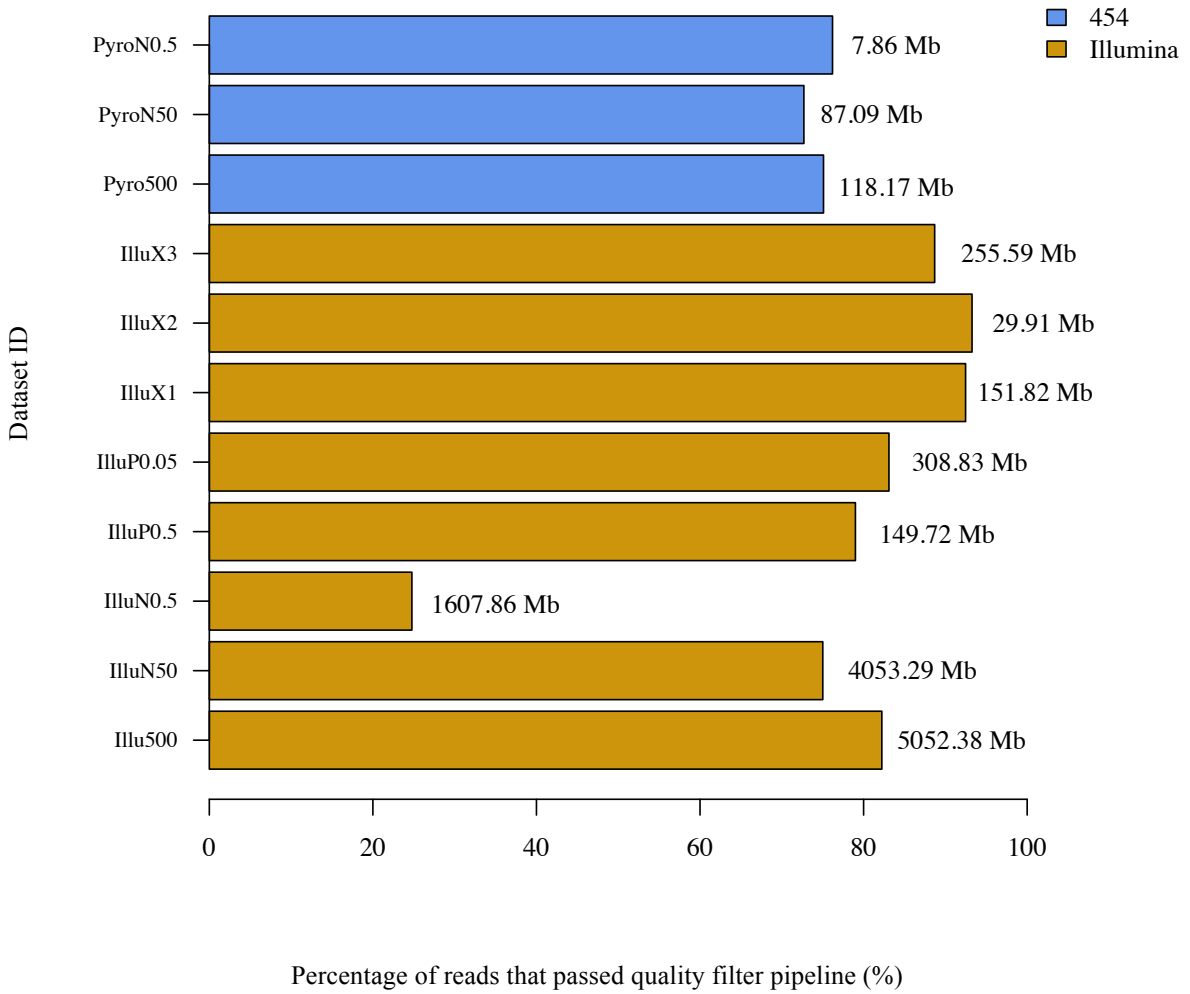


Figure 2.2. The proportion of high quality reads in each dataset. Bars represent the percentage of each dataset that passed the rigorous quality-filtering pipeline, and adjacent values depict the size of the resulting dataset. The percentage is calculated with respect to the total number of reads in the un-filtered raw dataset.

2.3.4 Predicted species abundance from read mapping

Many metagenomic analysis tools assume that the ratio of reads mapping to different microbial clades within a dataset is proportionate to their relative abundance within a microbial community. Therefore, as an equal number of nucleic acid molecules per species were used to create the synthetic microbial community, without considering any sequence composition or human error biases, it was assumed there should be an equal frequency of reads originating from each of the eleven genomes. Also as DNA rather than cellular material was used to construct the community there should be no chromosomal copy number bias towards any of the species. Therefore to determine if the relative abundance of reads mapping to each species was biased by the low level of input DNA utilised to generate some of the sequenced libraries, all filtered reads were mapped to the eleven reference genomes and the

number of uniquely mapping reads was transformed into relative abundance by normalising against the total filtered read count (Table 2.4).

Between 85% and 90% of each dataset aligned successfully against the eleven reference genomes (Table 2.4). As the synthetic microbial community comprised closely related species, the level of non-unique alignments was expected to be substantial, however less than 7% of the total mapped reads generated alignments to multiple genomic locations (Table 2.4). The relative abundance of reads mapping to each reference genome was significantly different from the expected distribution in the majority of datasets (X_2 test, p values < 0.05), with only the following two datasets showing no significant difference: IlluX1 and PyroN50 (X_2 test, all p-values > 0.05), indicating that read abundance did not accurately predict the relative abundance of each species (Fig 2.3).

To understand how accurately each individual species was represented within each dataset, and to investigate the impact of GC content, the rule was applied that if relative abundance was more or less than 25% away from the expected abundance of a particular species, it was classified as either 'under-represented' or 'over-represented' respectively, and if it was within the 25% range it was classified as 'correctly-represented'. The low GC bacteria *F. magna* was the only species to be consistently underrepresented in every dataset with a relative abundance ranging from 1.7% to 4% (Figure 2.3). Similarly, the low GC bacteria *S. lugdunensis* was the only species to be over-represented in every dataset with a maximum relative abundance of 26%, almost three times the expected relative abundance (Figure 2.3). As low GC content species were found to be over and under represented within datasets, and only a weak negative correlation was found between GC content and relative abundance, it is unlikely that GC content influences relative abundance (Pearson's product moment correlation test: r = - 0.64, p-values < 0.05).

When the read abundance distribution of the Illumina dataset generated using the standard library preparation technique (Tru-Seq) was compared to each of the Illumina datasets generated using low-input techniques (Nextera, Nextera XT, Parkinson), only the Nextera XT datasets exhibited a significantly different proportion of reads mapping to each reference genome (X_2 test, all p values < 0.001). It is possible that the contrasting digestion and amplification techniques utilised in Nextera XT and Tru-Seq protocols affected the proportion of DNA fragments originating from each genome in the final library. As the Nextera XT replicate one dataset exhibited a read abundance distribution similar to expected,

it is possible that this technique generates libraries with more accurate representation of the original DNA sample, however as the other Nextera XT replicates did not demonstrate this, more samples would be required for validation.

Table 2.4. The number of mapped and uniquely mapped reads aligning against the set of eleven reference genomes. As the Roche GS Mapper does not allow for non-unique read alignments, the number of mapped reads for all 454 datasets reflects the number of unique alignments. The percentage of mapped and uniquely mapped reads is relative to the total number of filtered reads.

Dataset Name	Mapped Reads		Uniquely Mapped Reads	
	No.	%	No.	%
Illu500	43,875,432	87.1	42,087,212	84.1
IlluN50	34,791,266	86.7	33,764,823	84.1
IlluN0.5	14,371,563	90.3	13,826,847	86.9
IlluP0.5	1,259,267	85	1,156,096	78
IlluP0.05	2,604,866	85.2	2,404,018	78.6
IlluX1	1,302,487	86.7	1,232,874	82
IlluX2	262,955	88.8	246,714	83.3
IlluX3	2,203,570	87.1	2,059,481	81.4
Pyro500	280,606	87.2	280,606	87.2
PyroN50	305,953	88.2	305,953	88.2
PyroN0.5	35,006	88.6	35,006	88.6

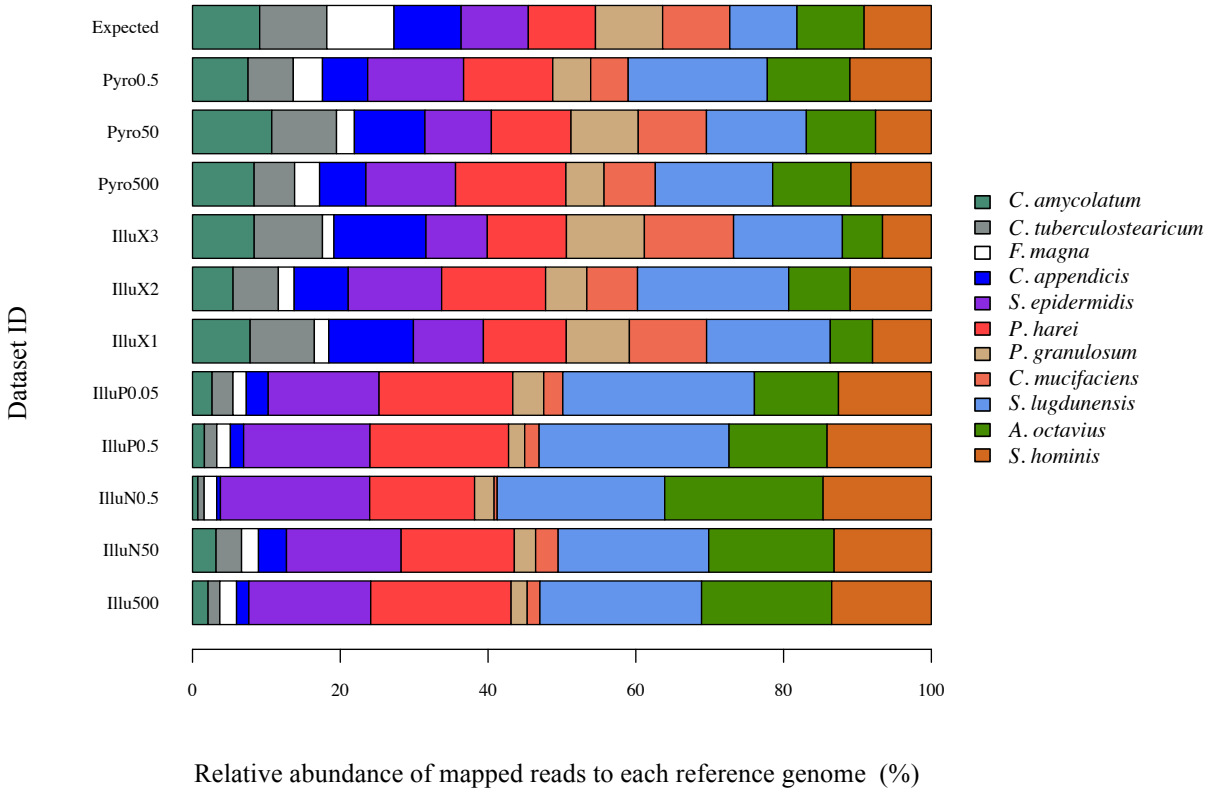


Figure 2.3. Relative abundance of the eleven species in each dataset verses the expected abundance calculated from read mapping counts normalised against total number of filtered reads

2.3.5 Phylogenetic reconstruction of metagenomic datasets

The taxonomic analysis of authentic metagenomic datasets cannot take advantage of reference genome sequences to estimate relative taxonomic abundances. Therefore to evaluate how accurately two popular non-reference-based taxonomic analysis tools could predict microbial composition, all datasets were analysed using two publically available tools, MetaPhlAn and MG-RAST, which both utilise different approaches to generate taxonomic profiles. MetaPhlAn employs an alignment-based approach utilising BLASTn to map reads against a database of clade-specific marker genes [198]. MetaPhlAn estimates relative abundance, therefore it was possible to directly compare datasets of differing sizes, removing the need to subsample and normalise the larger datasets, which may have resulted in

substantial biasing of the subsequent community composition predictions. MG-RAST is a web-based metagenomic analysis server which generates phylogenetic and functional profiles via a BLAT similarity search against the M5 non-redundant database [192]. To allow comparisons to be made between datasets, the number of taxonomically classified sequence feature hits was normalised by the total number of annotated features relative to each dataset.

MG-RAST and MetaPhlAn both over-estimated the microbial diversity of the synthetic microbial community, predicting an average of 152 and 56 species-level microbial clades per dataset, respectively. For both tools the level of predicted microbial diversity was significantly correlated with the size of the input dataset, suggesting that to a certain degree the number of species found may be an artefact of the number of reads (Pearson's product moment correlation test: $r = 0.92$ (MetaPhlAn) and $r = 0.82$ (MG-RAST), all p-values < 0.001). To remove extremely low abundance classifications, which may be sequence-related artefacts, relative abundance cut-offs of 0.1% and 0.5% were applied to all datasets. Applying a cut-off of 0.1% considerably reduced the false-diversity in all datasets, and resulted in a more accurate estimation of microbial diversity by both tools, with MetaPhlAn predicting an average of 20 species per dataset and MG-RAST predicting an average of 22 species per dataset (Table 2.5). When a cut-off of 0.5% was applied, MetaPhlAn and MG-RAST predicted the correct number of species in two and four datasets respectively, and the average number of predicted species dropped to 11 for MetaPhlAn and 12 for MG-RAST (Table 2.5). When genus-level microbial diversity was examined, MG-RAST out-performed MetaPhlAn by predicting the correct number of genera in five datasets at a cut-off of 0.1% and in all datasets at a cut-off of 0.5% (Table 2.5).

In comparison to the level of microbial diversity predicted in the Illu500 dataset, the number of classified species at a relative abundance cut-off of 0.5% was the same or higher for all reduced yield Illumina datasets, apart from the IlluN0.5 dataset, in which only one fewer species was detected (Table 2.5). This suggests that using a reduced level of microbial DNA does not diminish the level of microbial diversity captured when a relative abundance cut-off is applied (Table 2.5). Although cut-offs are not generally applied to taxonomic abundance estimations of real microbial communities, the substantial number of false-positive classifications generated with no cut-off suggests that a relative-abundance based filtering step may improve the accuracy of taxonomic composition estimations when examining real microbial communities.

As both classification methods were based on homology searches against public databases, the four members of the synthetic microbial community without publically available reference genomes were unsurprisingly not represented in the taxonomic predictions of any dataset using either tool. MG-RAST correctly estimated the presence of the remaining seven species at a relative abundance of 0.1% or over, however MetaPhlAn was unable to classify *C. tuberculostearicum* in any dataset. This indicates that *C. tuberculostearicum* is not represented in the MetaPhlAn dataset, which could either be a result of a technical error in the development of the database, or could indicate that the clade-specific marker genes selected for this species are not discriminatory enough.

The MG-RAST average microbial diversity estimate was over three times higher than MetaPhlAns prediction: therefore it could be argued that MetaPhlAn generates a more accurate representation of the true diversity of the synthetic microbial community. As MG-RAST utilises a combination of multiple databases including the NCBI nr, KEGG, COG, SEED, RDP, SILVA and Greengenes to generate taxonomic classifications, in comparison to the single highly discriminatory clade-specific marker-gene database employed by MetaPhlAn, it is possible that MG-RAST generated more false-positive hits simply due to the increased exposure to the possibility of incorrectly annotated proteins or ribosomal sequences within the databases [192,198]. Due to the uniqueness of the clade-specific marker genes utilised by MetaPhlAn it is claimed that this method is less sensitive to erroneous reads that may have been missed by the filtering step, which may also explain the different levels of microbial diversity estimated by MG-RAST and MetaPhlAn.

Table 2.5. The number of species-level and genus-level microbial clades detected by MetaPhlAn and MG-RAST in each dataset at relative abundance levels of 0.1% or over and 0.5% or over. 'MG' refers to MG-RAST, 'MP' refers to MetaPhlAn.

Dataset ID	≥ 0.1% Abundance				≥ 0.5% abundance			
	Species-Level		Genus-Level		Species-Level		Genus-Level	
	MG	MP	MG	MP	MG	MP	MG	MP
Illu500	21	19	7	12	10	9	6	6
IlluN50	25	22	6	14	16	9	6	7
IlluN0.5	19	18	9	10	12	8	6	5
IlluP0.5	23	19	7	13	16	10	6	7
IlluP0.05	19	22	6	14	11	10	6	9
IlluX1	19	21	6	15	12	13	6	8
IlluX2	29	18	9	13	14	11	6	9
IlluX3	20	28	9	18	11	15	6	12
Pyro500	23	19	7	10	13	11	6	7
PyroN50	23	21	6	13	11	13	6	9
PyroN0.5	18	9	6	7	11	9	6	7

2.3.6 Metagenomic assembly of Illumina datasets

Assembling single read data into longer contiguous sequences greatly enhances the yield of information that can be extracted regarding gene content, functional potential and phylogenetic classification. All filtered Illumina datasets were assembled using two *de Bruijn* graph based assemblers optimised for the assembly of metagenomic data: IDBA-UD, an algorithm that assumes uneven sequencing depths due to varying species abundances and MetaVelvet, an extension of the single-genome assembler Velvet [182,183]. A comparison of these two tools has not previously been reported.

The N50 lengths for all datasets varied between 300 bp and 24,000 bp, depending on the volume of data used for assembly (Pearson's product-moment correlation test: all p-values <

0.001). In comparison to IDBA-UD, MetaVelvet generated longer assemblies for the majority of datasets, with a larger N50 length in six out of eight datasets and a longer total contig length in five out of eight datasets (Table 2.6). For the majority of datasets MetaVelvet was also able to incorporate a larger proportion of reads into the final assembly than IDBA-UD. When the distribution of contig lengths was examined, the majority of contigs generated by both tools were between 101 bp and 1000 bp in length (Table 2.7). Due to the strong correlation identified between dataset size and N50 length it was not possible to generate inter-sample comparisons.

Table 2.6. Basic assembly statistics of all Illumina datasets assembled with IDBA-UD and MetaVelvet. 'I-U' refers to IDBA-UD and 'MV' refers to MetaVelvet. Percentage is in relation to the total number of either reads or contigs.

Dataset Name	N50 Length (bp)		No. of Contigs		Total Contig Length (mbp)		Assembled Reads (%)	
	I-U	MV	I-U	MV	I-U	MV	I-U	MV
Illu500	11,454	24,367	19,269	4,194	24.2	21.5	93.18	87.36
IlluN50	15,555	21,412	10,923	5,342	24.4	23	96.35	83.93
IlluN0.5	1688	2,526	14,098	13,229	12.4	13.7	82.62	84.79
IlluP0.5	438	645	12,408	15,086	5.3	7.7	49.00	67.71
IlluP0.05	756	1,051	17,146	13,836	10.7	10.6	75.91	77.48
IlluX1	467	445	21,481	40,029	9.7	15.2	59.38	85.36
IlluX2	325	247	514	3,718	0.2	1	6.45	18.17
IlluX3	838	5,129	24,540	9,681	15.8	18.2	80.94	91.05

Table 2.7. Contig length distributions for all datasets assembled with MetaVelvet and IDBA-UD. 'I-U' refers to IDBA-UD and 'MV' refers to MetaVelvet.

Dataset ID	Contig Length Distribution (% of total contigs)					
	101-1000 bp		1001-10000 bp		≥10001 bp	
	I-U	MV	I-U	MV	I-U	MV
Illu500	57.30	47.73	15.26	41.63	1.79	10.63
IlluN50	48.53	49.78	27.72	42.23	3.90	7.99
IlluN0.5	68.59	74.20	21.56	24.99	0.51	0.81
IlluP0.5	97.03	90.22	2.06	9.78	0.00	0.00
IlluP0.05	83.03	77.08	14.32	22.91	0.00	0.01
IlluX1	93.26	94.81	4.91	5.19	0.00	0.00
IlluX2	95.53	99.22	2.53	0.78	0.00	0.00
IlluX3	80.96	59.28	16.29	37.39	0.01	3.33

2.3.6.1 Understanding the accuracy of the assembled contigs

Although the length of an assembly is an important statistic used to determine the successfulness of an assembly, it is also imperative to understand the accuracy of the assembled contigs by examining how precisely they represent the reference genomes. This is particularly relevant for the assembly of metagenomic datasets as microbial communities generally consist of a multitude of closely related strains and subspecies sharing homogenous regions, which may result in the co-assembly of reads originating from different genomes, leading to the formation of chimeric contigs. Contig accuracy was examined by understanding the proportion of chimeric contigs present in the assembly and determining the degree of chimericity (Fig 2.4), [226,227]. A contig was defined chimeric if it contained uniquely mapping reads originating from more than one reference genome [226,227].

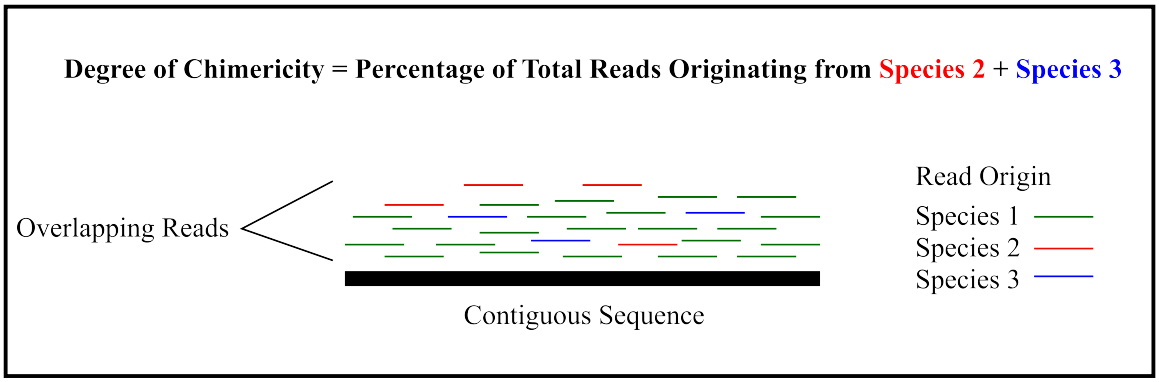


Figure 2.4. Overview of the method for generating the 'Degree of Chimericity' (DOC) statistic. The depicted contig is dominated by reads originating from species one, with a low number of reads originating from species two and three. The DOC represents the proportion of total reads that do not originate from the most dominant species, therefore in this situation the DOC is the proportion of total reads originating from species two and three.

2.3.6.2 Identifying chimeric contigs

The proportion of contigs comprising reads originating from different taxonomic groups was relatively high for all assemblies generated by both MetaVelvet and IDBA-UD, with an average of 39% and 57% of contigs classified as chimeric, respectively (Fig 2.5). Seven out of eight assemblies produced by MetaVelvet contained a lower proportion of chimeric contigs than their corresponding IDBA-UD assembly, signifying that MetaVelvet may be more adept at solving the correct path when repeats or homogenous sequences cause branches to form within the *de Bruijn* graph (Fig 2.5). Correspondingly MetaVelvet also managed to assemble a considerably

higher proportion of non-chimeric contigs for all datasets in comparison to IDBA-UD, and for four datasets over 50% of contigs were classified as non-chimeric (Fig 2.5).

The majority of the chimeric contigs generated by both tools exhibited a low degree of chimericity (DOC), with 14 out of 16 assemblies exhibiting an average DOC of below 14% (Fig 2.5). A low DOC signifies that contigs were assembled from a large number of reads originating from one species only, and only a low proportion of reads originated from a different taxonomic group. This statistic revealed that although a large percentage of contigs were classified as chimeric due to the co-assembly of reads originating from different species, the non-dominant species did not contribute a large proportion of the reads. Interestingly, the IDBA-UD assemblies of datasets Illu500 and IlluN50 exhibited average DOC scores of 45% and 48% respectively, which were between 3 and 12 times higher than DOC scores for all other datasets, including the corresponding MetaVelvet assemblies (Fig 2.5). It was determined that when IDBA-UD was utilised for assembly, the DOC of chimeric contigs exhibited a significant positive correlation with input dataset size that was not seen when MetaVelvet was used (Pearson's product-moment correlation test: $r = 0.94$, $p < 0.001$). This strong correlation explains the high DOCs observed in the Illu500 and IlluN50 datasets, which comprised an average of 4 gb of sequence data in comparison to the average 417 mb in all other datasets, and suggests that the algorithm employed by IDBA-UD may not deal with extremely large datasets as well as MetaVelvet. It has been previously shown contig length is positively correlated with DOC, however in all datasets assembled in this study, contig length was not correlated with DOC (Pearson's product-moment correlation test: all p values < 0.05) [227].

This high level of observed interspecies co-assembly within these datasets may have resulted from the presence of closely related species within the synthetic microbial community that share common regions of sequence that could not differentiated. The apparent ability of MetaVelvet to generate assemblies containing a lower proportion of chimeric contigs than IDBA-UD may be associated with the different k-mer strategies utilised by both approaches. Although the assembly strategies share certain similarities in that they both involve a graph decomposition step based on local k-mer coverage, a fundamental difference is the use of an iterative k-mer step by IDBA-UD, which begins with a small k-mer length and then uses resolved paths as reads for use with longer k-mer lengths, in comparison to the use of a single k-mer length by MetaVelvet [182,183]. It is possible the use of shorter k-mer lengths by IDBA-UD allows a larger number of interspecies overlaps to form during the initial *de Bruijn* graph iterations, which are not subsequently identified as erroneous prior to the next iteration. It is also possible that IDBA-UD is not as sensitive to erroneous k-mers originating from incorrect reads as MetaVelvet, which if not filtered out may lead to the formation of incorrect overlaps

and subsequently the production of chimeric contigs. This may also explain the association between dataset size and DOC observed in IDBA-UD chimeric contigs, as larger datasets are likely to contain a higher number of erroneous reads.

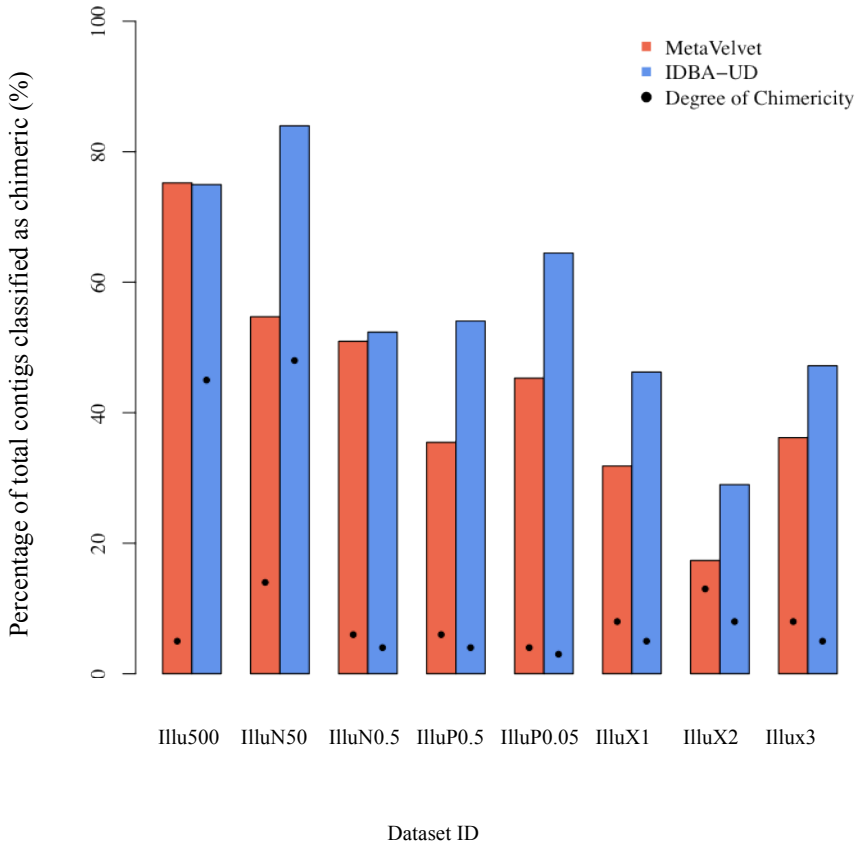


Figure 2.5. The proportion of chimeric contigs identified in the MetaVelvet and IDBA-UD assemblies of all Illumina datasets. The average degree of chimericity (DOC) is shown as a percentage of total reads mapping to a contig that did not originate from the dominant species within that contig (as described in Figure 2.5).

2.3.7 Metagenomic assembly of 454 datasets

All 454 datasets were assembled using the Roche GS *De Novo* Assembler, which was developed for specific use with 454 generated reads and is currently regarded as the most effective tool for *de-novo* 454 assembly, using single genome or metagenomic data. The Pyro500 and PyroN50 datasets comprised a considerably larger volume of data than the PyroN0.5 dataset, and subsequently generated much longer assemblies (Table 2.8). To examine contig accuracy the proportion of chimeric and non-chimeric contigs was calculated as previously described (Table 2.8). The majority of contigs generated by all three datasets were classified as non-chimeric, exhibiting a substantially lower level of chimeric contig formation

in comparison to the Illumina assemblies. The increased ability to resolve more accurate contigs is likely to be attributed to the longer read length of 454 generated datasets, which was an average of 270 bp in comparison to the 101 bp generated by Illumina. Also the Roche GS *De Novo* Assembler employs a Overlap-Layout-Consensus assembly which carries out a pair-wise comparison of all reads, which is obviously a much too computationally expensive to apply to Illumina datasets, but generates a considerably more accurate assembly.

Table 2.8. Basic assembly statistics for all 454 datasets assembled using the Roche GS *De Novo* Assembly tool. Percentages were calculated with respect to either the total number of reads or total number of contigs.

Dataset ID	Total Bases (mb)	N50 Length (bp)	Total Contig Length (mbp)	Assembled Reads (%)	Chimeric Contigs (%)	Non-Chimeric Contigs (%)
Pyro500	125.9	2,404	18.3	88.29	21.57	68.66
PyroN50	100.3	1,507	16.7	81.29	19.03	73.52
PyroN0.5	8.9	539	0.1	5.69	26.32	50.72

2.3.8 Annotation of Illumina assembled contigs

All MetaVelvet and IDBA-UD assembled contigs were subject to feature prediction and functional annotation using the online metagenomic analysis server IMG/M [218]. Annotations were extracted in the form of Clusters of Orthologous protein Groups (COGs) to allow subsequent comparison between datasets and reference genomes. To understand how comprehensively each dataset represented the functional potential of the synthetic microbial community, the resulting collection of COG groups for each dataset was compared to the expected repertoire of COG groups generated from the set of eleven reference genomes. The term ‘false-negative’ was used to describe any entry annotated in the reference genomes but not the datasets and the term ‘false-positive’ was used to describe any entry annotated in the dataset but absent from the reference genomes. The abundance or number of alignments made to each identified COG group was also directly compared between the reference genomes and datasets, and entries were classified as either ‘over-abundant’ or ‘under-abundant’, and to allow direct comparison between datasets numbers were normalised using the total COG group count for each dataset. This abundance was then correlated to the expected abundance obtained from the IMG/ER annotation of the reference genomes and Pearson’s correlation coefficients were calculated for each comparison.

Both IDBA-UD and MetaVelvet assembled contigs generated a relatively low level of false-positive COG alignments, with an average 10% of total COG groups classified as false-positives using both tools (Table 2.9, Figure 2.6). Although this accounts for only a small proportion of the total functional diversity, it represents a considerable number of incorrectly annotated genes that would be impossible to differentiate from accurate annotations in a genuine metagenomic analysis. It is also likely that a small proportion of the false-positive classifications are a result of the draft genome status of the majority of the reference sequences. Understanding the level of false-negative annotations is also a good indication of how comprehensively the functional diversity of the microbial community has been represented by each dataset, as it describes the number of COG groups annotated in the reference genomes that are essentially 'missing' from the metagenomes. In all datasets a strong negative correlation was observed between the total contig length and the number of false-negative COG groups, indicating unsurprisingly that the more sequence data utilised to generate the annotations, the fewer COG groups are missed (Pearson's product-moment correlation test: all r scores < -0.7, all p values < 0.05). In seven out of eight Illumina datasets the MetaVelvet assembled contigs exhibited a lower number of false negative annotations than the corresponding IDBA-UD assembled contigs, indicating that contigs generated by MetaVelvet were able to capture a larger proportion of the functional diversity (Table 2.9, Figure 2.6). This is likely to be a result of the increased N50 lengths and total contig lengths of the MetaVelvet assemblies.

To judge the overall accuracy of the functional profiles generated by IMG/M using MetaVelvet and IDBA-UD assembled contigs, the number of genes annotated to each COG group was calculated for each dataset and compared to the expected abundance generated from annotation of the reference genomes (Table 2.9, Figure 2.6). MetaVelvet contigs were able to generate a slightly more accurate representation of functional abundance of the synthetic microbial community than the corresponding IDBA-UD contigs, assigning the correct number of annotations to an average of 17.6% of COG groups, in comparison to an average of 14.1% using IDBA-UD contigs (Table 2.9, Figure 2.6). A relatively high proportion of COGs were characterised as over- abundant and under-abundant in contigs generated by both MetaVelvet and IDBA-UD, indicating that although the correct functions and genes were identified, the relative abundance of the majority of functions was predicted incorrectly. Relative abundance is frequently utilised as a discriminating feature to identify important genes, however the inaccuracy at which relative abundance has been predicted in these datasets indicates that it may not reflect the true abundance within the microbial community, and therefore may lead to the generation of false conclusions.

Correlations between the expected and observed COG abundances in each dataset revealed that the MetaVelvet assembly of the IlluN50 dataset generated the most accurate functional representation of the microbial community (Pearson's product-moment correlation test: $r = 0.94$, p value < 0.05). Apart from the IlluX2 dataset, all Illumina datasets generated good correlations to the expected COG abundances and there was limited deviation between the correlations of MetaVelvet and IDBA-UD assembled contigs of corresponding datasets, showing comparable functional profiles generated by contigs from both tools (Pearson's product-moment correlation test: all r scores > 0.8, all p value values < 0.05).

As the majority of MetaVelvet assemblies generated a more comprehensive representation of the COG-based functional diversity of the synthetic microbial community, it could be concluded that MetaVelvet contigs were ultimately more accurate.

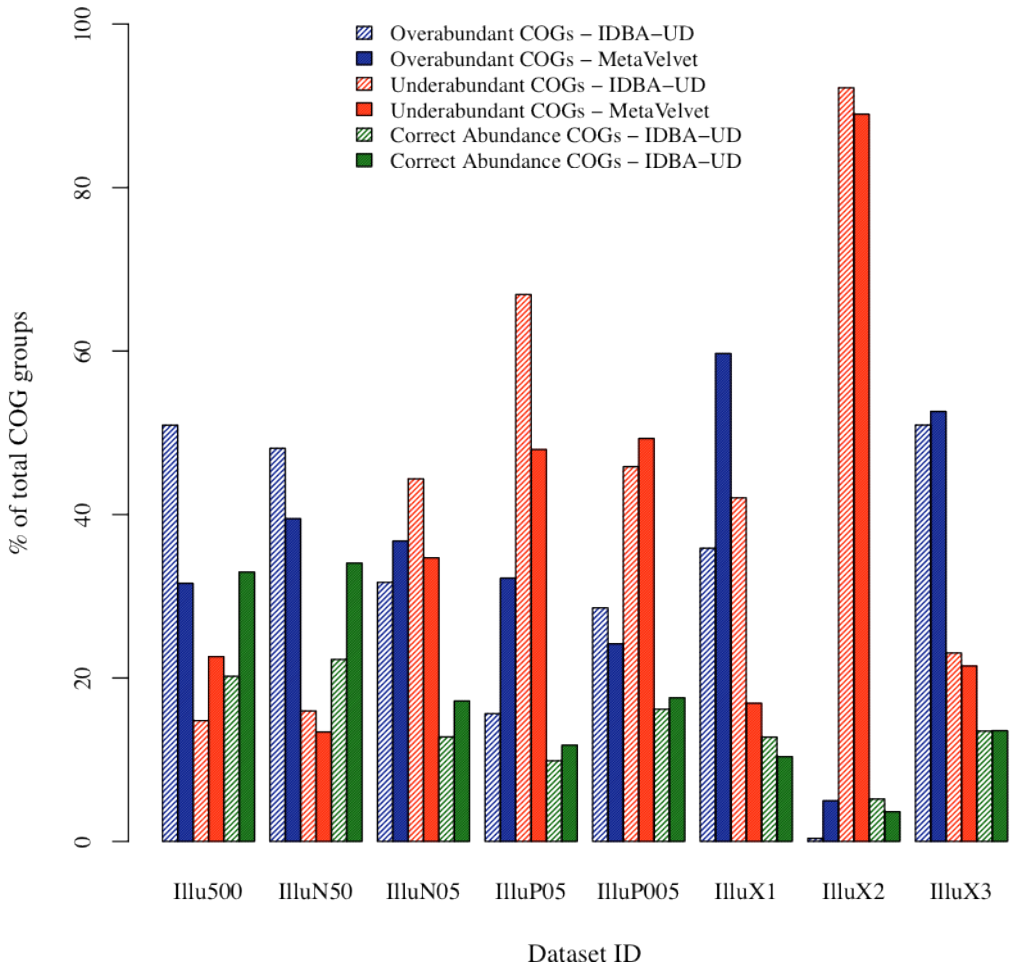


Figure 2.6. The proportion of overrepresented, unrepresented and correctly represented COG groups annotated by IMG/M in each MetaVelvet and IDBA-UD assembled dataset with respect to the expected COG abundance found in the reference genomes.

Table 2.9. The number of false-negative COG groups and the proportion of false-positive COG groups annotated by IMG/M within each MetaVelvet and IDBA-UD assembled dataset with respect to the expected COG abundance found in the reference genomes. The percentage was calculated in relation to the total number of COG groups annotated to each dataset. As the false-negative COG groups represent annotations missing from the datasets, they could not be transformed into a percentage and are therefore listed as raw counts.

Dataset Name	False-negative (raw count)		False-positive (%)	
	I-U	MV	I-U	MV
Illu500	104	53	12.85	14.07
IlluN50	104	47	13.05	13.84
IlluN0.5	416	291	11.36	11.12
IlluP0.5	607	455	8.04	7.58
IlluP0.05	277	304	8.95	9.37
IlluX1	322	174	13.02	9.31
IlluX2	2055	1300	2.39	2.22
IlluX3	201	165	12.37	12.48

2.3.9 Annotation of unassembled metagenomic data

To understand how accurately the functional profile was represented by unassembled metagenomic data, a selection of datasets were annotated using the EBI Metagenomic Service, which unlike the IMG/M pipeline, allows feature prediction and functional annotation of unassembled reads [257]. The effect of read length was also examined by utilising both short-read Illumina datasets and longer-read 454 datasets. In collaboration with the InterPro team at the EMBL-EBI site in Hinxton, Cambridge, the unassembled Illu500, IlluN50, IlluN0.5, Pyro500 and PyroN50 datasets were functionally annotated using InterProScan v5.0, which generates annotations in the form of InterPro (IP) entries, which represent a specific protein signature including protein families, domains, binding sites, active sites, repeats or post-translational modifications (PTM) [257]. The resulting collection of IP hits for each dataset was compared to the expected repertoire of IP matches generated from the set of eleven reference genomes, and the number of under-abundant, over-abundant, correctly abundant, false-positive and false-negative hits were calculated and normalised based on total IP count. Although it is not possible to directly compare the number of InterPro entries and COG groups, if the total number of COG groups/IP entries defines the complete functional diversity of the community,

it is subsequently possible to determine how comprehensively assembled and unassembled reads represent this.

A relatively low proportion of total IP entries were classified as false-positives, indicating that utilising unassembled reads does not lead to a high level of inaccurate alignments, which may have been expected due to the short fragment length (Table 2.10, Figure 2.7). Almost 1,000 false-negative IP entries were classified per dataset, indicating that a considerable proportion of the functional diversity was not accessible using unassembled reads. Almost double the number of false-negative entries were generated by the Illumina datasets in comparison to 454 datasets, indicating the 454 datasets generated a much more comprehensive coverage of the functional profile of the microbial community (Table 2.10, Figure 2.7). Although the Illumina datasets contained on average 35 times more base pairs of sequence data than the 454 datasets, it is apparent that the large volume of data did not compensate for reduction in read length, which is obviously an essential feature for functional prediction. The current trend in metagenomic studies to employ shorter read-length platforms in the pursuit of large datasets with high coverage may result in the neglect of a large proportion of the functional potential of a microbial community.

To assess how accurately each dataset predicted functional abundance, the number of hits generated against each IP entry by the reference genomes and the metagenomes was compared. Illumina datasets generated a lower average number of under-abundant IP entries in comparison to 454 datasets, indicating that the increased data volume allowed a more comprehensive representation of certain IP entries, however when abundance data was subject to correlation, the 454 datasets generated the strongest positive correlations (Pearson's product-moment correlation test: $r = 0.89$ and 0.82, all p values < 0.001, Fig 5).

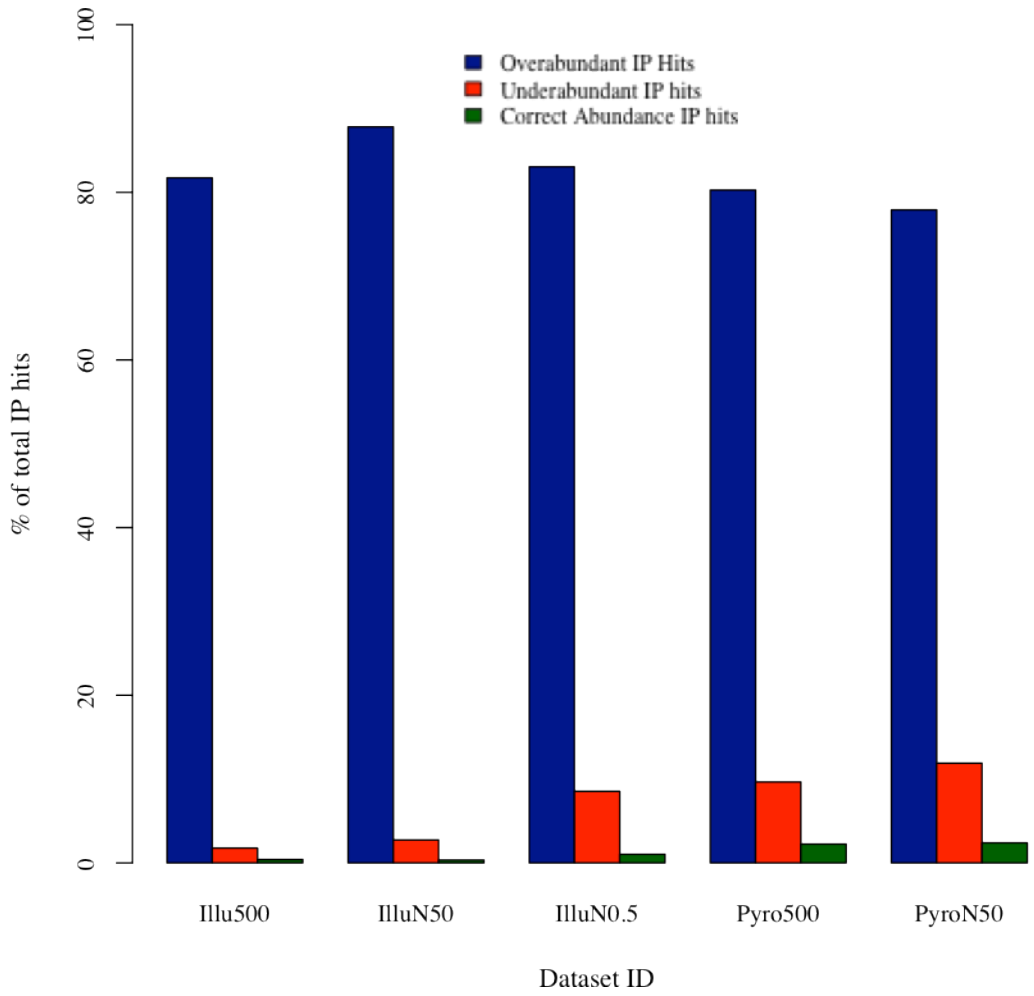


Figure 2.7. The proportion of overrepresented, unrepresented and correctly represented InterPro entries in three Illumina and two 454 unassembled datasets with respect to the expected InterPro entry abundance found in the reference genomes.

Table 2.10. The proportion of false-positive and the number of false-negative InterPro entries. The percentage is calculated in relation to the total number of InterPro entries annotated to each dataset, and as the false-negative hits represent InterPro entries absent from the dataset, the abundance cannot be transformed into a percentage.

Dataset Name	False-positive (%)	False-negative (raw count)
Illu500	16.12	1147
IlluN50	9.12	1169
IlluN0.5	7.39	1280
Pyro500	7.83	673
PyroN50	7.81	726

2.4 Conclusion

To ensure the success of any metagenomic analysis project, it is essential to understand which of the available bioinformatics tools will generate the most accurate representation of the taxonomic and functional content of a microbial community. To interrogate the accuracy of a number of popular computational tools, multiple datasets were generated from an *in vitro* synthetic microbial community, and subject to comparative metagenomic analysis using a range of metagenomic analysis programs. Many tools have previously been validated using *in silico* metagenomic data, however this is the first application of an *in vitro* community for comparative analysis of multiple metagenomic tools. This is also the first report of generating Nextera, Nextera XT and Parkinson based sequencing libraries from a bacterial community.

It is assumed that the proportion of reads originating from each taxonomic group within a metagenomic dataset represents the relative abundance of that particular clade within a microbial community. However, the majority of datasets exhibited significantly different read abundance distributions than expected. This suggests that read distribution is not an accurate estimator of microbial relative abundance within a microbial community, and it is possible that sequence composition differences between species may have led to the over or under representation of certain microbial clades. Therefore, it is apparent that using total read counts to estimate microbial relative abundance will result in inaccurate predictions of taxonomic composition.

When comparing the accuracy of two popular taxonomic prediction tools, MG-RAST and MetaPhlAn it was found that the latter generated a more accurate prediction of the taxonomic

profile of the microbial community. MG-RAST determined taxonomic composition by calculating the proportion of total reads that generated significant alignments against sequences of known taxonomic origin, in comparison to the marker-gene based approach utilised by MetaPhlAn. Therefore, MG-RAST is sensitive to incorrect read distributions within the dataset, which may have arise as a result of genome composition differences, as previously discussed. This may explain the more accurate profile generated by MetaPhlAn, which only generates alignments against a small proportion of each genome, and is therefore less sensitive to read distribution errors. As the taxonomic composition of a genuine metagenomic dataset is completely unknown, it is often challenging to distinguish which taxonomic classifications are artefacts of the library preparation or sequencing protocol, and which are authentic. This highlights a major limitation of utilising a metagenomic approach to profile a microbial community, and suggests that species diversity will always be overestimated.

The accurate assembly of metagenomic data into longer contiguous sequences is essential for the comprehensive prediction of the functional potential of a microbial community. This work demonstrated that contigs generated by both MetaVelvet and IDBA-UD represented a considerable proportion of the microbial functional diversity of the microbial community. A slightly more accurate functional profile was predicted by MetaVelvet assembled contigs, presumably as a consequence of the generation of longer and less chimeric contigs in comparison to IDBA-UD. It is possible that the incorporation of an iterative k-mer strategy by IDBA-UD results in the inclusion of erroneous k-mers in the initial stages of the assembly, which may introduce branches into the *de Bruijn* graph leading to the production of shorter and more inaccurate contigs.

This chapter presented a thorough validation of a number of popular metagenomic analysis tools using an *in vitro* simulated microbial community of known composition. Deducing which tool generated the most faithful prediction of the taxonomic and functional composition of a microbial community will allow future metagenomics studies to generate the most accurate analyses possible. This work also revealed that it is not possible to generate a completely accurate reconstruction of a microbial community using a metagenomics approach, and reinforced the fact that caution is required when interpreting predictions of taxonomic and functional diversity from genuine microbial samples. Although an additional aim of this chapter was to understand if it was possible to generate accurate metagenomic data from low-input library preparation kits such as Nextera, Nextera XT and the method developed by Parkinson *et al*., the extreme variation in dataset size between samples prevented a statistically significant comparison between high and low yield datasets [240].

2.5 Recommendations for future metagenomic analysis projects

- Taxonomic composition should be estimated using a marker-gene based method such as MetaPhlan, rather than a method that relies on total read count, in order to negate the effects of inaccurate read distributions due to unknown clade-specific biases.
- To remove possible false-positive taxonomic classifications it is advisable to apply a relative-abundance cut-off of at least 0.1% to predicted taxonomic profiles.
- Assembled contigs were able to represent a considerable proportion of the functional diversity of the microbial community; therefore it is recommended that short-read metagenomic datasets should be assembled prior to functional annotation.
- When assembling short-read metagenomic datasets MetaVelvet should be used in preference of IDBA-UD to ensure the most accurate contigs are generated.

CHAPTER 3

Metagenomic analysis of the taxonomic composition and gene content of the axillary microbiota in relation to axillary malodour

3.1 Introduction

3.1.1 Structure and topography of the axilla

Due to its distinctive topographical characteristics the axilla boasts a warm, moist and sheltered environment that supports the growth of a dense microbial community [29]. The axilla is host to a high density of apocrine, eccrine and sebaceous glands that provide a plentiful supply of secreted odourless fluids comprised of proteins, lipids, steroids and cholesterol [262]. Continuous secretions from sweat and sebaceous glands maintain the moist environment of the axilla and provide a constant supply of nutrients for the inhabiting microbiota. Eccrine sweat glands are distributed throughout the entire skin surface and continuously secrete a watery-like fluid, whilst apocrine sweat glands do not become active until puberty, secrete a low level of nutrient rich fluid comprising proteins, steroids, lipids and electrolytes and are restricted to body regions associated with hair, such as the axilla, mammary and genital areas [6,262,263]. Each sweat gland responds to a different stimulus, with apocrine glands linked to emotional stimuli such as pain or anxiety and eccrine glands exhibiting a thermoregulatory role, and therefore responding to temperature changes [5,263-265]. Apoeccrine glands have been proposed as a third type of sweat gland which consistently secrete a fluid similar in composition to eccrine sweat, however their existence is currently in contention [4,5,266].

3.1.2 Microbial composition of the axilla

Research into the taxonomic composition of the axillary microbiota began in the 1960's when it was first realised that volatile compounds could be generated from odourless apocrine secretions upon interaction with cutaneous microorganisms [267]. In 1981 Leyden *et al*. examined the axillary microbiota of 229 participants using culture-dependent techniques and determined that the three genera, *Staphylococcus, Corynebacterium* and *Propionibacterium* dominated the majority of axillary communities [29]. Taylor *et al*. employed a similar technique, including the application of 16S rRNA gene sequencing to a subset of corynebacterial isolates, to characterise the microflora of 61 axillary samples [268]. This study determined that the axillary microflora was

dominated by either *Staphylococcus* spp., aerobic coryneforms or less commonly *Propionibacterium* spp., and found a significant association between high axillary malodour and total levels of aerobes and aerobic coryneforms [268]. 16S rRNA gene sequencing and classification of a subset of axillary *Corynebacterium* spp. isolates by Taylor *et al*. revealed matches to only two sequences in the RDP database at that time, *Corynebacterium* sp. G-2 CDC G5840 and *C. mucifaciens* DMMZ 2278, with the former subsequently re-classified as *C. tuberculostearicum* [268,269]. A subsequent study by Grice *et al.* which characterised the microbial diversity at 20 skin sites generated over 100,000 near full-length 16S rRNA gene sequences via cloning and Sanger sequencing and determined that axillary microbial communities were dominated by either *Staphylococcus, Corynebacterium* or *Betaproteobacteria* spp., agreeing with the previous study by Taylor *et al.* [52]. They also uncovered a high degree of interpersonal variation with respect to the microbial community composition, which has been repeatedly observed in subsequent studies [22,52]. The first completely culture-independent characterisation of the skin utilised direct sequencing of the 16S rRNA gene, and determined from a pooled sample that *Staphylococcus* spp. dominated the axillary microbiota, with substantial proportions of *Corynebacterium* and *Propionibacterium* spp. also present [50]. In 2010 Gao *et al*. employed an alternative culture-independent method to enumerate bacterial and fungal numbers at multiple body sites including the axilla [270]. Using 16S and 18S rRNA qPCR it was determined that the axilla was host to the highest number of bacterial 16S rRNA gene copies, indicating a dense microbial population [270]. Species belonging to the *Corynebacterium, Propionibacterium* and *Staphylococcus* genera were again found to be prevalent within the majority of axillary samples, however substantial proportions of *Streptococcus* spp. were also identified [270]. A recent study by Egert *et al*. determined that the majority of axillary bacteria are metabolically active, after finding similar relative levels of corresponding rDNA and rRNA sequences [22]. Egert *et al*. again identified *Corynebacterium, Propionibacterium* and *Staphylococcus* as prevalent axillary genera, and also found a substantial abundance of *Anaerococcus* and *Peptoniphilus* in the majority of axillary samples with a higher abundance of the latter associated with the right axilla, suggesting a relation between handedness and *Peptoniphilus* levels [22]. In a study published this year Callewaert *et al*. characterised the axillary microbiota of 53 subjects using denaturing gradient gel electrophoresis (DGGE) and determined that people are either dominated by *Staphylococcus epidermidis* (61%) or one of *Corynebacterium* spp., *Proteobacteria* or *Staphylococcus hominis* (39%) [271]. They observed the temporal stability of the axillary microbiota and found it to be relatively stable, although in two subjects there was a shift in microbial dominance from *Corynebacterium* to *Staphylococcus* and from *Corynebacterium* to *Proteobacteria* following a change in climate and an increase in physical activity respectively [271]. Nine samples underwent 16S rRNA sequencing from which they

confirmed that *Staphylococcus* and *Corynebacterium* species dominate the majority of people, with over half dominated by *Staphylococcus* [271].

3.1.3 Biotransformation of steroids

The biotransformation of odourless steroid precursors by the axillary microbiota is thought to be one of the main routes towards generating axillary malodour, however the biochemical pathways mapping out the enzymatic reactions and precursors involved have yet to be fully understood. It is thought that malodorous steroids attribute their characteristic to the orientation of functional groups in relation to the steroid nucleus, as it has been shown that certain 3α-sterols, whose C3 hydroxyl group is below the plane have more intense odours than their counterpart 3β-sterols, whose hydroxyl group is above the plane [77,78]. The main source of steroidal malodour in the axilla is thought to be attributable to 16-androstenes, which due to their absence in freshly produced apocrine sweat are hypothesised to originate from an odourless steroid precursor [263,267]. The most abundant 16-androstenes detected in malodorous sweat include 5α-androst-16-en-3-one (androstene), 5α-androst-16-en-3α-ol (androstenol) and 4,16- androstadien-3-one (androstadienone), which have all been linked to urine and musk like odours [264]. It was evident in early research that certain aerobic corynebacteria possess the enzymatic ability to generate androstene and androstenol from unknown precursors in apocrine secretions [272]. In more recent work Decreau *et al*. identified the odourless steroid androsta-5,16-dien-3β-ol as a major precursor of malodorous 16-androstene steroids, including the highly malodorous steroid androsta-4,16-dien-3-one (androstadienone), which is prevalent in apocrine sweat [78,273]. They predicted a novel steroidal biotransformation pathway leading from odourless precursors to malodorous C16-androstene steroids, and determined that an interplay between the aerobic corynebacterial enzymes 4,5- or 5α reductase, 3α(β)-sterol dehydrogenase and steroid 4,5 isomerase was required for successful generation of malodorous 16-androstene steroids [78]. Other work has also characterised the enzymatic ability of aerobic corynebacteria to generate malodorous 16-androstenes, and it has been repeatedly shown that 16-androstenes can only be generated via the biotransformation of precursors already containing the C16 double bond [74,78]. It was also found that generation of 16-androstenes requires the combination of multiple aerobic corynebacteria, and recently Austin *et al.* reported that no single aerobic corynebacterial isolate was able to catalyse the full complement of biotransformation reactions carried out by the collection of mixed isolates, and only a very low number of axillary isolated isolates were capable of C16-steroid biotransformation at all [74]. Although 16-androstene steroids are undoubtedly present in axillary sweat, the lack of recent evidence directly implicating 16-androstenes in axillary malodour, and the fact that an estimated 50% of the

population demonstrate anosmia to 16-androstenes has caused dispute regarding the extent of their contribution towards axillary malodour [273,274].

3.1.4 Degradation of long chain fatty acids

The axilla is home to a myriad of fatty acids originating from apocrine, eccrine and sebaceous secretions and many studies have reported that short and medium chain branched volatile fatty acids (VFAs) ($C_2 - C_{11}$) are responsible for a large proportion of axillary malodour [264,273]. A number of metabolic pathways have been identified in corynebacteria, staphylococci, propionibacteria, micrococci and brevibacteria that are capable of generating VFAs from specific substrates.

VFA generation was initially attributed to the partial catabolism of long chain ($C_{14} - C_{30}$) methyl-branched fatty acids by corynebacterial β-oxidation activity, leading to the production of shorter chain volatile metabolites such as the abundant axillary odorant isovaleric acid (3-methylbutyric) [73]. It was observed that an abundance of corynebacteria exhibiting a 'lipophilic' phenotype, defined by an *in vitro* dependence upon the addition of exogenous fatty acids, was correlated with the production of malodorous VFAs [79,272]. The lipophilic phenotype, which is caused by the lack of a fatty acid synthase (*fas*) gene responsible for the biogenesis of lipids, is often confused with the ability of certain corynebacterial species to catabolise fatty acids, however not all lipophilic corynebacteria are able to carry out this biochemical function and certain non-lipophilic strains do exhibit lipid-catabolising activity [73,79,272]. Therefore it was proposed that lipid-catabolising corynebacteria were the primary source of VFA generation in axillary secretions. Although the corynebacterial genes required for β-oxidation have not been biochemically characterised, a recent bioinformatic study found that lipophilic species *C. jeikeium, C. urealyticum* and *C. kroppenstedtii* contained the full repertoire of *fadA, fadB, fadE* and *fadH* homologs whilst non-lipophilic species *C. diphtheriae, C. aurimucosum* and *C. glutamicum* lacked certain genes [80,275]. Recent work has questioned the extent to which this corynebacterial mechanism contributes towards axillary VFA generation due to the scarcity of axillary-isolated corynebacteria with the ability to catabolise lipids, and the lack of appropriate VFA precursors in sebaceous secretions [268,269].

Propionibacteria and staphylococci have also been implicated in the generation of VFAs via the metabolism of common skin substrates such as glycerol, lactic acid and branched aliphatic amino acids [80]. Glycerol is liberated via the hydrolysis of the triglyceride component of both apocrine and sebaceous secretions, and lactic acid is an abundant skin compound [276]. Both are fermented by staphylococci and propionibacteria to generate the short chain VFAs propionic

(C_3) and acetic (C_2) acid, which are potential components of axillary malodour [80]. Staphylococci also possess the enzymatic ability to convert branched aliphatic amino acids such as valine, leucine and isoleucine into malodorous short chain fatty acids (C_4-C_6) [80]. It is unclear whether axillary propionibacteria generate VFAs through the fermentation of amino acids, as the enzymatic ability has only been reported in *Propionibacterium* spp. not found in the axilla [80,277]. Axillary amino acids originate from both eccrine secretions and via the actions of bacterial proteases acting upon proteins originating from the keratinising epidermidis, and are highly abundant within this specific niche. Due to the substrate abundance, and high relative abundance of staphylococcal species within the axillary microbiota, it is now thought that the main source of short-chain VFA generation is via staphylococcal metabolism of amino acids into highly malodorous fatty acids, although no correlation has been found between axillary malodour levels and staphylococcal relative abundance within the axillary microbiota [268,269].

3.1.5 Nα-acylglutamine aminoacylase (N-AGA) cleaved volatile fatty acids

In addition to the short and medium chain VFAs previously described, a significant proportion of axillary malodour is thought to be attributable to a group of structurally unusual medium-chain fatty acids, mainly (E)-3-methyl-3-hexenoic acid (3M2H) and its hydrated form (RS)-3-hydroxy-3-methylhexanoic acid (HMHA) [72,73,278-280]. In early research 3M2H was initially thought to be secreted non-covalently bound to apocrine secreted binding proteins 1 and 2 (ASOB1 and ASOB2), with ASOB2 later identified as the high-density apolipoprotein D (ApoD), belonging to a family of lipid transporting carrier proteins [281,282]. However there are conflicting hypothesis regarding the composition of the bound conjugate with more recent work revealing that 3M2H and HMHA are actually secreted covalently bound to L-glutamine residues which are subsequently released by a corynebacterial zinc dependent Nα-acylglutamine aminoacylase (N-AGA) [72,278,279]. In 2009 Troccaz *et al*. identified substantial amounts of a HMHA-Gln conjugate in axillary sweat indicating the presence of glutamine bound precursors [283]. However, more recently the presence of a covalently bound HMHA-ApoD precursor was characterised in axillary sweat but there was no evidence of an association between 3M2H and ApoD [284]. This led the author to hypothesis that HMHA-ApoD may in fact be a precursor of HMHA-Gln, releasing the glutamine bound conjugate following cleavage by an uncharacterised endopeptidase, however that theory is unlikely due to the high levels of ApoD required to be present to allow the production of the reported concentrations of HMHA present in axillary sweat [284,285].

The corynebacterial aminoacylase N-AGA is proposed to play a central role in axillary malodour generation, with the enzymatic ability to release a wide range of axillary odorants from their associated glutamine conjugates, including the highly prevalent odorant HMHA [72]. The enzyme is encoded by the corynebacterial gene *agaA*, which was initially characterised and cloned from *Corynebacterium* strain Ax20 [269,278]. The enzymatic activity is very specific for the glutamine residue, and is unable to liberate axillary odorants conjugated to alternative amino acids including asparagine, aspartate and glutamate [278]. Natsch *et al*. found that N-AGA was able to liberate at least 26 fatty acids from their suspected glutamine conjugates when incubated with fresh axillary sweat, highlighting its essential role in axillary malodour formulation [72].

3.1.6 Sulphur-containing compounds

Volatile sulphur compounds (VSCs) have a significant influence upon axillary malodour due their low olfactory threshold, and have been characterised as emitting meaty, onion or fruity type odours. The most prevalent thioalcohol in axillary sweat is 3-methyl-3-sulfanylhexan-1-ol (3M3SH), however numerous others have also been identified including 2-methyl-3-sulphanylbutan-1-ol(2M3SB), 3-sulphanylpentan-1-ol and 3-sulphanylhexan-1-ol [75,76,88]. Gly-Cys-(S)-conjugates of 3M3SH, 3-sulphanylhexan-1-ol and 2-methyl-3-sulphanyl-pentan-1-ol were identified in fresh axillary sweat, indicating that volatile thiols are secreted as Gly-Cys bound precursors [286]. Previously, Natsch *et al*. had determined that a corynebacterial C-S lyase was able to cleave a synthetic Cys-(S)-3M3SH conjugate releasing the malodorous thiol and that staphylococcal isolates were unable to catalyse this reaction [88]. It was subsequently determined that the action of two corynebacterial enzymes, a C-S β-lyase and a metal dependent dipeptidase, encoded by corynebacterial genes *aecD* and *tdpA* respectively, were required to release 3M3SH from its Gly-Cys-(S)-conjugate precursor [287]. *Corynebacterium* strain Ax20 and *C. jeikeium* were both found to exhibit this enzymatic ability, and although C-S β-lyase activity was not initially detected in staphylococcal strains, an axillary isolated *S. haemolyticus* Ax4 was able to generate malodorous sulphur compounds from fresh axillary sweat, suggesting a possible role of staphylococci in malodour generation [76].

3.1.7 The first genetic link to malodour

A significant advance towards our understanding of the genetic control of body odour was the recent characterisation of a single nucleotide polymorphism (SNP: 538G>A) in the gene ABCC11, which was correlated with a reduction in axillary odour [288]. The ABCC11 gene encodes an ABC-driven efflux pump that localises to the axillary apocrine glands and is thought to be involved in the transport and/or biosynthesis of glutamine bound axillary

odorants. People who were homozygous for this SNP, which changes the side chain from glycine to arginine, produced significantly lower amounts of amino-acid conjugates of axillary odorants including HMHA-Gln, 3M2H-Gln and Cys-Gly-(S)-3M3SH, than people who were heterozygous for the SNP or possessed the wild-type genotype [288]. The AA genotype is much more prevalent in Asians who are commonly associated with very low body odour in comparison to Caucasians and Africans, in which the heterozygote and wild type genotypes are more prevalent [288].

3.1.8 Aims of the chapter

Due to the low biomass of skin-associated samples, characterisation of the axillary microbiota has been limited to 16S rRNA-based techniques. To generate a comprehensive description of the taxonomic composition of the axillary microbiota, and to further elucidate the role of the axillary microbiota in malodour generation, this chapter presents the first application of metagenomic sequencing to axillary samples isolated from high and low malodour environments.

3.2 Methods

3.2.1 Axillary sampling and malodour assessment

Ethical clearance for this study was obtained by Unilever Research and Development, Port Sunlight, who subsequently isolated and supplied all axillary samples (Unilever R&D, Port Sunlight, HOL 11.262). Samples were isolated from the left and right axillary vaults of 10 healthy male volunteers using a slightly altered version of the Williamson-Kligman cup-scrub technique [289]. Volunteers were provided with a control ethanolic deodorant and soap and advised to use only these products in their axillae for two weeks prior to the study. Sampling included placement of a sterile Teflon cup (9.62 cm^2) in the axillary vault followed by the application of 2.55 ml of sampling buffer (50 mM Tris-HCL + 0.1% v/v Triton-X100, pH 7.9). Using a sterile Teflon stick the surface of the skin was gently agitated for 1 min and the sampling buffer was collected. This process was repeated to yield a total of 5.10 ml sample buffer per axilla. Axillary wash samples were aspirated into sterile tubes and stored at -20°C until required.

The left and right axillary vaults of the ten subjects were each subject to malodour assessment by a Unilever team of six expert odour assessors. Odour intensity was assigned a score on the scale of 0-5, with nought representing no malodour and five representing an intense malodour. Malodour intensity scores from all six expert assessors were averaged to generate the mean malodour score (MMS) (Appendix A Table 1).

3.2.2 Extraction of whole genomic DNA

Prior to DNA extraction, axillary samples were defrosted on ice and cells were collected via centrifugation. Cells were then re-suspended in 200-500 µl TE buffer (pH 8.0) and transferred to Pathogen Lysis L tubes (QIAGEN) where up to an additional 100 µl of TE buffer was added to equalise the sample volumes. To each tube, 4 µl of Ready-Lyse Lysozyme (250 U μl^{-1}) (EPICENTRE) was added and the sample was incubated for 18 h at 37°C with gentle shaking (300 rpm). After incubation bead beating was carried out on a FastPrep cell disruptor for 40 s at a speed of 6 m/s, then repeated with samples held on ice for 5 min in-between (MP BIOMEDICALS). DNA was extracted with the automated QIAsymphony DNA extraction robot using the Qiagen Virus/Pathogen DNA extraction kit and the ComplexFix 800 protocol with carrier RNA substituted for sterile molecular grade water (QIAGEN). DNA was eluted in 60 µl AVE elution buffer and the three samples per organism were pooled. DNA was quantified

with the Qubit dsDNA high sensitivity (HS) kit with the Qubit 2.0 spectophotometer (INVITROGEN). Due to the low DNA yields achieved it was not possible to check for degradation on an agarose gel.

3.2.3 Nextera XT library preparation

Due to the extremely low DNA concentration of the extracted axillary samples (0.0016 ng μl^{-1} to 1.09 ng μl^{-1}) it was necessary to construct the libraries using the transposition based library preparation kit Nextera XT, which allows library preparation from 1 ng total gDNA (ILLUMINA). For each of the 20 axillary samples as close to 1 ng of gDNA as possible was fragmented and tagged via action of the Nextera transposome complex, with specific adaptor sequences added for subsequent identification. Following tagmentation, the transposase enzyme was neutralised and purified, and the fragmented DNA was amplified via a limited-cycle PCR reaction. Final libraries were quantified with the Qubit dsDNA high sensitivity (HS) kit with the Qubit 2.0 spectophotometer (INVITROGEN) and 1 µl was run on an Agilent BioAnalyser machine with a high sensitivity chip (HS) (AGILENT). Due to the low concentration of a number of the libraries it was decided not to size-select the libraries. Libraries were pooled in eqimolar amounts and sequenced on two lanes of the Illumina HiSeq (ILLUMINA). The Centre for Genomic Research (CGR) at the University of Liverpool carried out the cluster generation and sequencing of all samples.

3.2.4 Bioinformatic Analysis

3.2.4.1 Pre-processing of raw data

All datasets underwent a strict filtering protocol to remove low-quality, human contaminating and artificially duplicated reads. The CGR employs a standard read-filtering pipeline on all sequenced datasets which comprised: *i)* the removal of Illumina adaptor sequences using Cutadapt v1.2.1: *ii*) the trimming of low-quality bases using Sickle v1.2, which utilises a sliding window of a defined size to remove read segments which do not have a minimum phred quality value of 20 and *iii*) the removal of any trimmed reads below 10bp in length ([290], https://github.com/najoshi/sickle).

The NCBI Best Match Tagger tool (BMTagger) identifies contaminating human sequences by comparing query 18mers against an indexed human genome reference database [291]. The human genome build 37 (GCA_000001405.1) was downloaded from NCBIs FTP server and indexed

into a BLAST database using BMTagger, which was then used to remove all human contaminating reads from the axillary datasets. Artificial duplicates, introduced during the PCR step in the library preparation protocol, were identified and removed using PRINSEQ v0.20.3 using the exact duplicate parameter using paired-end data [292]. In contrast to the earlier version of PRINSEQ used to filter the synthetic datasets in chapter two, this newer version of PRINSEQ allowed paired-end duplicate removal.

3.2.4.2 Taxonomic profiling

The taxonomic composition of each sample was characterised using the previously described tool MetaPhlAn, based on its validation in chapter two [198].

Graphical representations of the taxonomic profiles including heatmaps and phylogenetic trees were generated using MetaPhlAn conversion scripts and the tree generator software GraPhlAn [198]. For subsequent statistical comparisons relative abundance predictions were transformed into OTU counts and formatted into the standard OTU BIOM format using a supplied MetaPhlAn script. β-diversity was assessed via calculation of the Bray-Curtis dissimilarity coefficient for all pair wise sample combinations using QIIME [294]. A coefficient of 0 indicates that the taxonomic profile of two samples are identical in terms of clade abundance and membership, whilst a score of 1 indicates a completely divergent taxonomic profile with no overlap [295]. Further statistical analysis was carried out using the Statistical Analysis of Metagenomic Profiles (STAMP) software [296]. Species with significant differences in relative abundance were identified using Fishers exact test with Story FDR multiple test correction, and biologically significant differences were extracted by applying a ratio of proportions (RP) effect size cut-off of ≥ 2 [296]. Rarefaction curves were generated using the community ecology package ‘vegan’ in the statistical programming language R, and comparative analyses were generated using MetagenASSIST [297].

3.2.4.3 LEfSe analysis of enriched microbial clades

Taxonomic abundance profiles from high and low odour samples were compared using a linear discriminant analysis (LDA) effect size (LEfSe) algorithm to identify microbial clades most likely to account for the difference in malodour levels [298]. LEfSe initially identified features that were differentially distributed between high and low odour samples using the non-parametric Kruskal-Wallis (KW) sum-rank test with a p-value cut-off of 0.05 [299]. Microbial clades which violated the null hypothesis of equal distributions were then subject to a pair wise Wilcoxon test which removes microbial clades which did not violate the null hypothesis for all pairs of high and low samples [298]. The differentially abundant set of features were then subject to LDA, which involved creating an LDA model using low and high malodour as dependent variables and taxonomic relative abundance as independent variables. The effect size of each differentially abundant microbial clade was subsequently calculated based on the average difference between the means of the high and low odour samples. The LDA effect size estimates the degree of responsibility associated with each microbial clade with respect to high or low axillary malodour. An LDA effect size cut-off of two was applied to identify clades exhibiting biologically significant differences between high and low odour samples.

3.2.4.4 Metagenomic assembly and annotation

All filtered reads were assembled into contiguous sequences using the de-Bruijn graph based assembly tool MetaVelvet, identified in the chapter two as the optimal metagenomic assembler in comparison to IDBA-UD [182,183].

All assembled contigs were then submitted to the Integrated Microbial Genomes Metagenomic (IMG/M) website for annotation [218]. The IMG/M gene finding protocol began with the identification of CRISPR elements using a combination of two programs, the CRISPR recognition tool (CRT) and PILER-CR [245,253]. Non-coding RNAs were then found using tRNAScan-SE-1.2 to identify tRNAs and internally developed rRNA models to identify rRNAs [218,244]. Protein-coding genes were discovered using a conjunction of four gene finding programs: *i)* GeneMark v.2.6r: *ii)* Metagene v.Aug08: *iii*) Prodigal and *iiii*) FragGeneScan [199,202,247,254]. Functional annotation began with a homology search against the Pfam-A database followed by an RPS-BLAST search against the database of position-specific scoring matrices (PSSM) for COGs using an e-value cut-off of 0.01 [205]. A similarity search against the non-redundant IMG database was then carried out to generate assignment of KEGG Orthology terms (KO) and enzyme commission (EC) numbers [214,218].

To allow comparison of functional profiles the number of genes classified to each COG was calculated for each dataset and normalised based on the total number of COG aligned genes. The previously described LDA tool LEfSe was used to identify COGs enriched within high and low axillary malodour samples.

3.3 Results and discussion

3.3.1 Axillary sampling and malodour assessment

Axillary samples were obtained from the left and right axillary vaults of ten subjects who will subsequently be referred to as P1-P10. Prior to sampling all subjects underwent axillary malodour assessment by trained assessors at Unilever R&D, Port Sunlight, using an internally approved technique that resulted in the assignment of a mean malodour score (MMS). MMS scores ≥ 3 were categorised as high, whilst scores < 3 were categorised as low. The left and right axillae of subjects P1, P2, P4 and P5 were categorised with low axillary malodour with MMS scores ranging from 1.5 to 2.67, whilst both axillae of subjects P7, P8 and P9 were categorised with high axillary malodour with MMS scores ranging from 3.17 to 3.67 (Table 3.1). Remaining subjects P3, P6 and P10 were categorised with one axilla emitting a high level axillary malodour and one axilla emitting a low level of axillary malodour (Table 3.1).

3.3.2 Generation of ultra-low concentration metagenomic sequencing libraries

The difficulty of extracting a sufficient yield of DNA is the major limiting factor preventing metagenomic analysis of a large number of skin-associated sites. Previous studies have resorted to combining samples from multiple sites and multiple subjects to generate the required DNA yields to create sequencing libraries, however this lowers the ability to perform intra- and inter-personal variation analyses [53]. This is the first study to perform whole-genome metagenomic analysis of the skin microbiota without either combining samples or performing a DNA amplification step prior to library generation. Following a robust extraction protocol, DNA was successfully extracted from all axillary samples with DNA concentrations ranging from 0.0008 ng μl^{-1} to 1.0859 ng μl^{-1} (Table 3.2). Illumina libraries were successfully generated from all samples using the Nextera XT low-input library preparation kit. Large high-quality datasets were generated for all samples apart from sample 7L, which was subsequently excluded from the study (Table 3.2). An average of 37 ± 5 million reads per dataset were generated for all remaining datasets.

Table 3.1. Malodour intensity scores of the left and right axillary vaults of ten male subjects. Assessments were carried out by six odour assessors at Unilever R&D, Port Sunlight

	Malodour Intensity Scores							
Sample ID	**Assessor 1**	**Assessor 2**	**Assessor 3**	**Assessor 4**	**Assessor 5**	**Assessor 6**	**Average score**	**Malodour level**
1L	1	3	2	2	1	2	1.83	Low
1R	2	2	4	3	2	1	2.33	Low
2L	2	2	2	3	3	2	2.33	Low
2R	3	3	3	2	2	3	2.67	Low
3L	1	2	1	2	2	3	1.83	Low
3R	3	4	3	3	4	2	3.17	High
4L	2	1	1	1	2	2	1.50	Low
4R	1	2	2	3	3	3	2.33	Low
5L	1	2	2	2	3	2	2.00	Low
5R	2	3	1	1	2	1	1.67	Low
6L	2	2	3	3	3	3	2.67	Low
6R	3	4	4	3	4	4	3.67	High
7L	4	4	5	3	3	3	3.67	High
7R	5	3	4	4	4	2	3.67	High
8L	5	5	5	5	4	4	4.67	High
8R	4	4	4	5	5	3	4.17	High
9L	3	3	3	3	4	3	3.17	High
9R	4	4	2	4	3	4	3.50	High
10L	2	2	2	2	2	4	2.33	Low
10R	4	3	3	3	3	3	3.17	High

Table 3.2. Malodour levels and sample information for the 19 axillary samples isolated from ten male subjects. Sample 7L was excluded from the study and is therefore not listed in this table.

Sample ID	Subject ID	Axillary Location	Malodour Level	Concentration of Extracted DNA (ng μl^{-1})
1L	P1	Left	Low	0.0016
1R	P1	Right	Low	0.0782
2L	P2	Left	Low	0.1291
2R	P2	Right	Low	0.1297
3L	P3	Left	Low	0.0009
3R	P3	Right	High	0.0008
4L	P4	Left	Low	0.0183
4R	P4	Right	Low	0.0475
5L	P5	Left	Low	0.0158
5R	P5	Right	Low	0.0418
6L	P6	Left	Low	0.0197
6R	P6	Right	High	0.0331
7R	P7	Right	High	0.2278
8L	P8	Left	High	0.0016
8R	P8	Right	High	0.0023
9L	P9	Left	High	0.3503
9R	P9	Right	High	0.1511
10L	P10	Left	Low	0.1113
10R	P10	Right	High	0.279

3.3.3 Removal of human originating sequence

As the sampling process involved vigorous agitation of the axillary surface it was likely that a considerable proportion of the resulting sample would comprise human originating cellular material, leading to contamination of the dataset with human reads. Therefore, BMTagger was used to identify and remove reads of human origin from each dataset. The proportion of human contaminating reads varied considerably between samples, ranging from 0.4% to 81% of the total number of reads in each dataset (Fig 3.1). This level of human contamination has also been observed in dental swab and plaque derived datasets [300]. Sample 3L predominantly comprised human originating reads, with over 27 million reads classified as contamination by BMTagger (Fig 3.1). Sample 1L also demonstrated a high level of human contamination with 60% of the original high-quality dataset classified as human contamination (Fig 3.1). On average, just over 24% of reads within each dataset were classified as human, however the standard deviation was 21% reflecting the wide spread of the data. A negative correlation was observed between the proportion of human classified reads and the DNA concentration of the axillary wash sample, indicating that samples with lower initial DNA concentrations were associated with a higher level of human sequence contamination in the final datasets (Pearson's product moment correlation test: $r = -0.56$, $p < 0.05$, Table 3.2). It is likely that the axillary samples exhibiting very low microbial DNA concentrations were not colonised by a high number of microbial cells, and therefore the sampling of these sites collected primarily human cells, leading to a significant amount of human-originating sequence in the final dataset.

It is known that artificial duplicates can bias the resulting taxonomic composition and abundance predictions of metagenomic datasets. Consequently, all duplicated reads thought to be artefacts of the library preparation protocol were removed using PRINSEQ v0.20.3 [301]. An average of 10% ± 4% of reads were identified as artificial duplicates per dataset and were removed (Fig 3.1). The proportion of artificially duplicated reads was quite consistent from dataset to dataset, indicating it is an artefact of the standardised library preparation protocol rather than the sample itself.

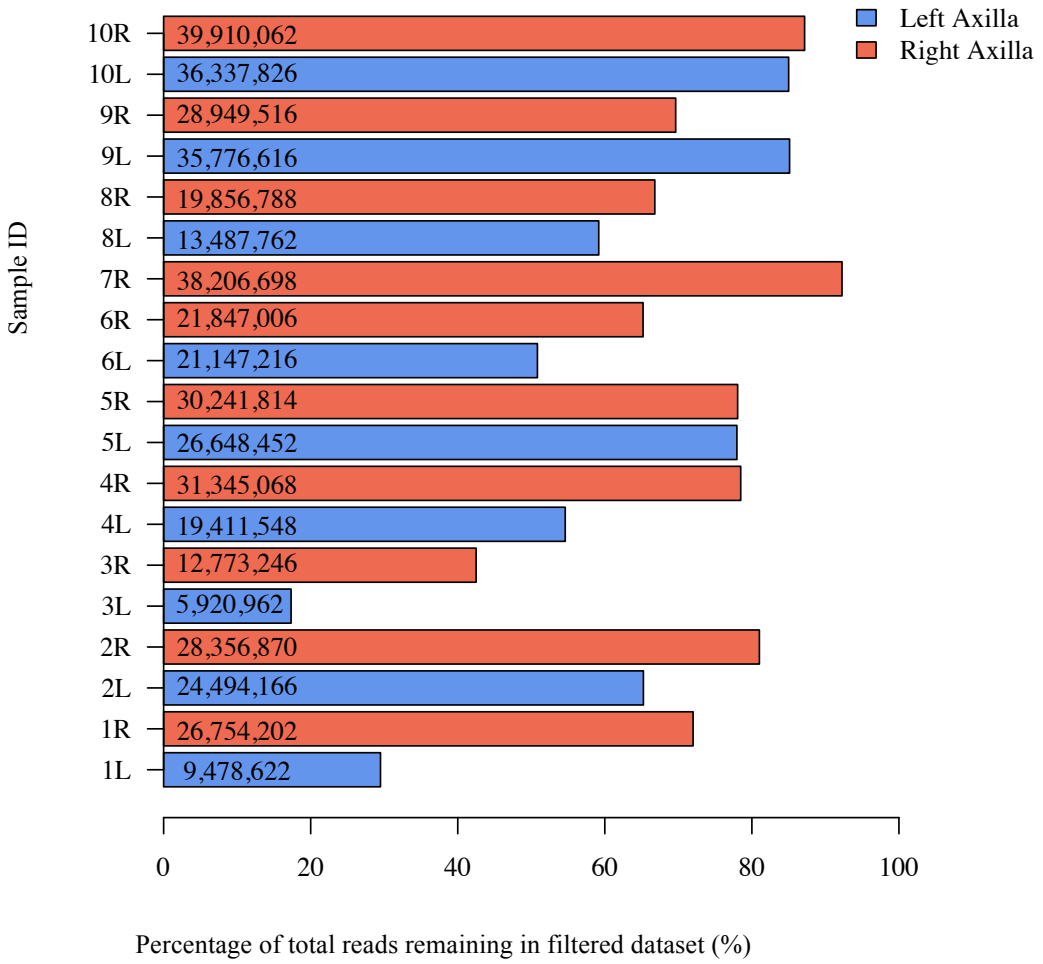


Figure 3.1. The proportion of original reads remaining in final dataset following filtering for human originating and artificially duplicated reads using BMTagger and PRINSEQ. The numbers within the bars represent the number of reads remaining in each filtered dataset.

3.3.4 Microbial composition of the axillary microbiota

The microbial composition of all filtered axillary datasets was predicted using MetaPhlAn, which utilises a clade-specific marker gene database to determine the taxonomic composition of the microbial community in terms of relative abundance [198]. The total length of all hits against each marker gene was normalised by the length of that specific marker gene, allowing calculation of the number of genomic copies of each taxonomic group, rather than the number of read hits against each taxa. Since MetaPhlAn estimates relative abundance, it was possible to directly compare the community composition of datasets of differing sizes, removing the need to subsample and normalise the larger datasets, which may have resulted in substantial biasing of the subsequent community composition predictions. Also, due to the clade specificity of the marker genes, it was possible to accurately assign taxonomic classifications to species level.

Although 35 phyla were detected in total, the axillary microbiota was dominated by *Firmicutes* and *Actinobacteria,* which on average accounted for over 99% of the relative abundance of each sample. The most abundant genus within the *Firmicutes* phylum was *Staphylococcus,* which was present at a relative abundance of 75% or higher in 15 out of 19 samples (Fig 3.2 and Fig 3.3). The *Actinobacteria* phylum was comprised of mainly *Propionibacterium* spp. and *Corynebacterium* spp., however *Propionibacterium* spp. were on average much more abundant than *Corynebacterium* spp. (Fig 3.2 and 3.3). The dominance of *Propionibacterium* over *Corynebacterium* has not been observed by previous studies, which usually reveal either a co-dominance or complete dominance by *Corynebacterium* spp. [22,268,270]. As 16S rRNA gene sequencing has a number of known associated biases including the preferential amplification of certain microbial clades based on primer set selection, it is possible that earlier studies have mis-estimated the true abundances of *Corynebacterium* and *Propionibacterium*, and that using a metagenomics approach, which is free of biases associated with the 16S rRNA gene sequencing method, has generated a more accurate representation of individual relative abundance [302].

Across all samples the axillary community was dominated by *S. epidermidis, S. hominis* and *P. acnes*, which dominated twelve, two and four samples respectively, and accounted for between 53% and 98% of the total microbial abundance in each sample (Fig 3.2 and Fig 3.3). Although *Corynebacterium* was identified as a dominant genus of the axillary microbiota in previous studies, no samples were dominated by this group of bacteria, and no corynebacterial species accounted for more than 6% of the total microbial abundance of any sample [21,271]. In sample 4R, *S. hominis* and *S. epidermidis* were both dominant species, accounting for 49% and 48% of total relative abundance respectively (Fig 3.2). It was found that in the majority of samples, over 95% of the species present contributed less than 10% of the total relative abundance, and in the one sample in which this rule did not apply, 99% of the species present accounted for only 19% of the total abundance (Fig 3.2). From the community composition of all sequenced axillary samples in this project, the typical structure of the axillary microbiota seemed to be defined by the presence of one dominant species that comprised at least 50% of the total community abundance, and a large number of very low-abundance species. The dominance by a low number of highly abundant taxa seems to be a signature of human-associated microbial communities, as this composition pattern has been observed in urogenital, skin, nasal, gastrointestinal and oral associated microbial samples [85].

To understand the most prevalent microbial clades within the axillary microbiome, the top 20 abundant species were identified in each sample. *S. epidermidis, S. capitis, S. hominis, P. acnes*

and at least one *Corynebacterium* spp. were present in every sample as one of the 20 most abundant species. Although no single *Corynebacterium* spp. was ubiquitous, *C. pseudogenitalium* was abundant in all but one sample, and *C. tuberculostearicum* was present as one of the top 20 most abundant species in 14 out of 19 samples. The species classification of the most abundant corynebacterial species in the axillary microbiota has until now been unknown, as current studies have utilised high-throughput direct 16S rRNA gene sequencing, which often allows classification to the genus level only [50,52,271]. Previous culture-dependent studies have reported the presence of *C. mucifaciens, C. afermentens, C. amycolatum, C. genitalium, C. riegelii, C. striatum* and *C. minutissimum* within the axillary microbiota, however due to the techniques employed, their relative abundance was unknown [268].

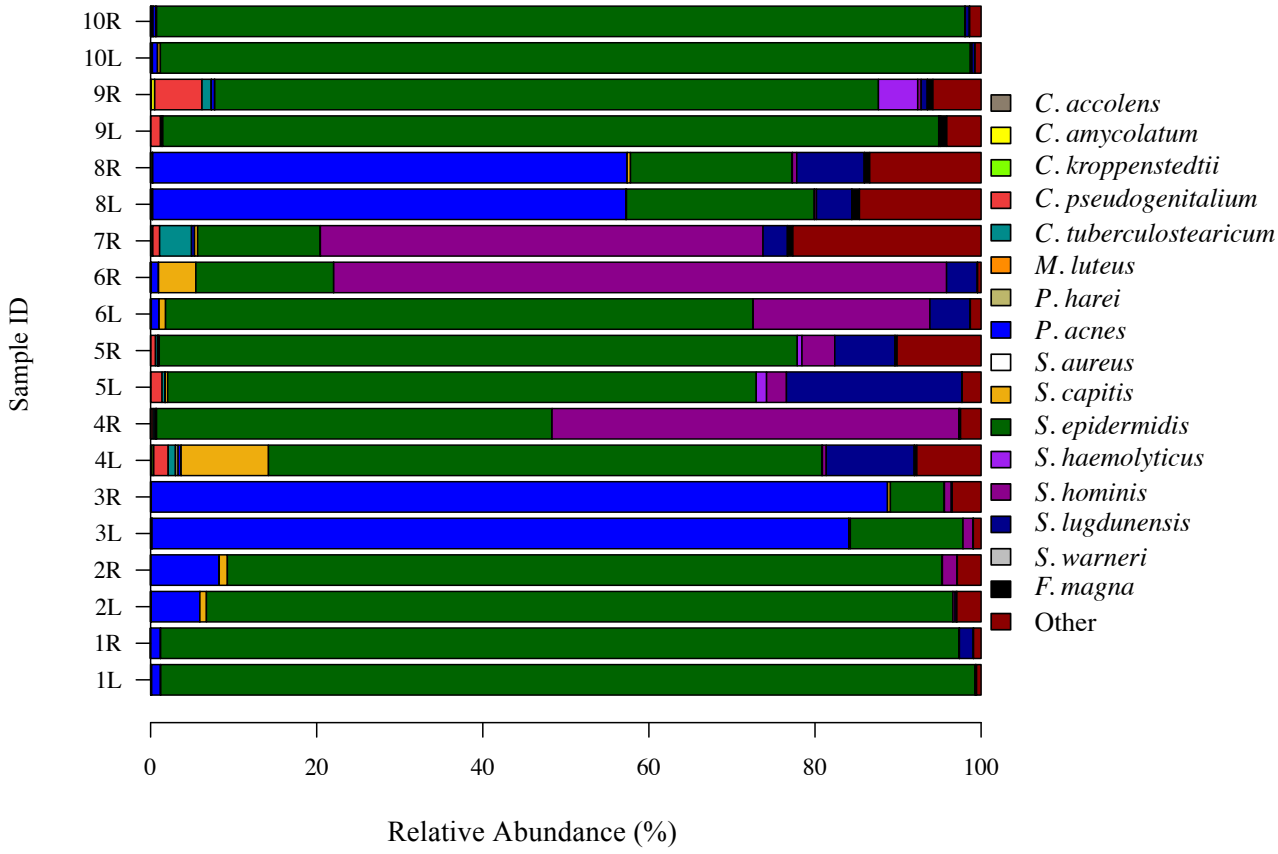


Figure 3.2. The taxonomic composition of the axillary microbiome. The relative abundance of the top 16 most abundant taxa are shown. Taxonomic profiles were generated using MetaPhlAn.

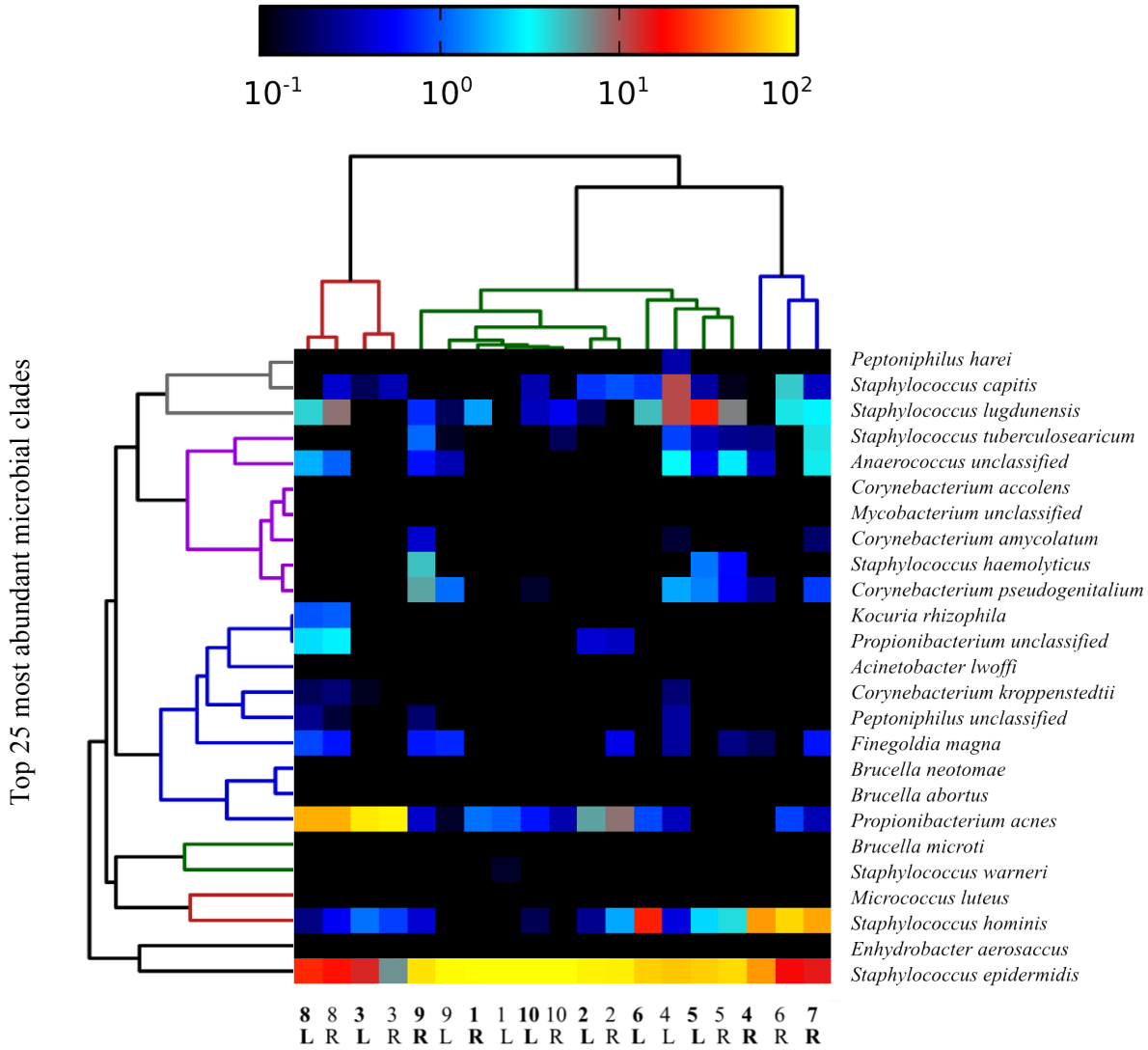


Figure 3.3. Representation of the top 25 most abundant clades in all samples. Hierarchical clustering of samples was performed using the Bray-Curtis method using MetaPhlAn. The black to yellow scale represents the relative abundance of each microbial clade in each dataset. Image was generated using a MetaPhlAn conversion script.

3.3.5 Inter-sample comparison of species richness

Species richness represents the number of distinct species within a microbial community, and does not take into account the abundance of each species, which is often referred to as species evenness.

Due to the stringent filtering pipeline imposed upon all datasets there was considerable variation in the number of high quality reads in each final dataset. To generate meaningful comparisons of species richness between datasets of different sizes, it was necessary to first

understand how comprehensively species diversity has been represented within each community. Therefore, the MetaPhlAn predicted taxonomic profiles of each sample were subject to rarefaction analysis, which calculates the average species richness represented by increasing numbers of individual microbes. Rarefaction curves generated for all datasets levelled off, indicating that additional sampling, or additional reads would not lead to the identification of any new species-level classifications, and therefore estimating that the species richness has been comprehensively represented by the current datasets (Fig 3.4).

Alpha diversity was shown to vary quite considerably between samples, with sample 7R predicted to be the most diverse with a species diversity of over 1,260 (Fig 3.4). The least diverse community was identified as sample 1L, which was predicted to contain just 341 species (Fig 3.4). With an average species richness of 906 ± 300, the diversity of the axilla is relatively low in comparison to other body sites such as the hand and the gut, which are both estimated to host approximately 3,000 species [303]. This reduction in diversity in comparison to other body sites is expected due to the sheltered and closed nature of the axilla, which reduces exposure to transient organisms.

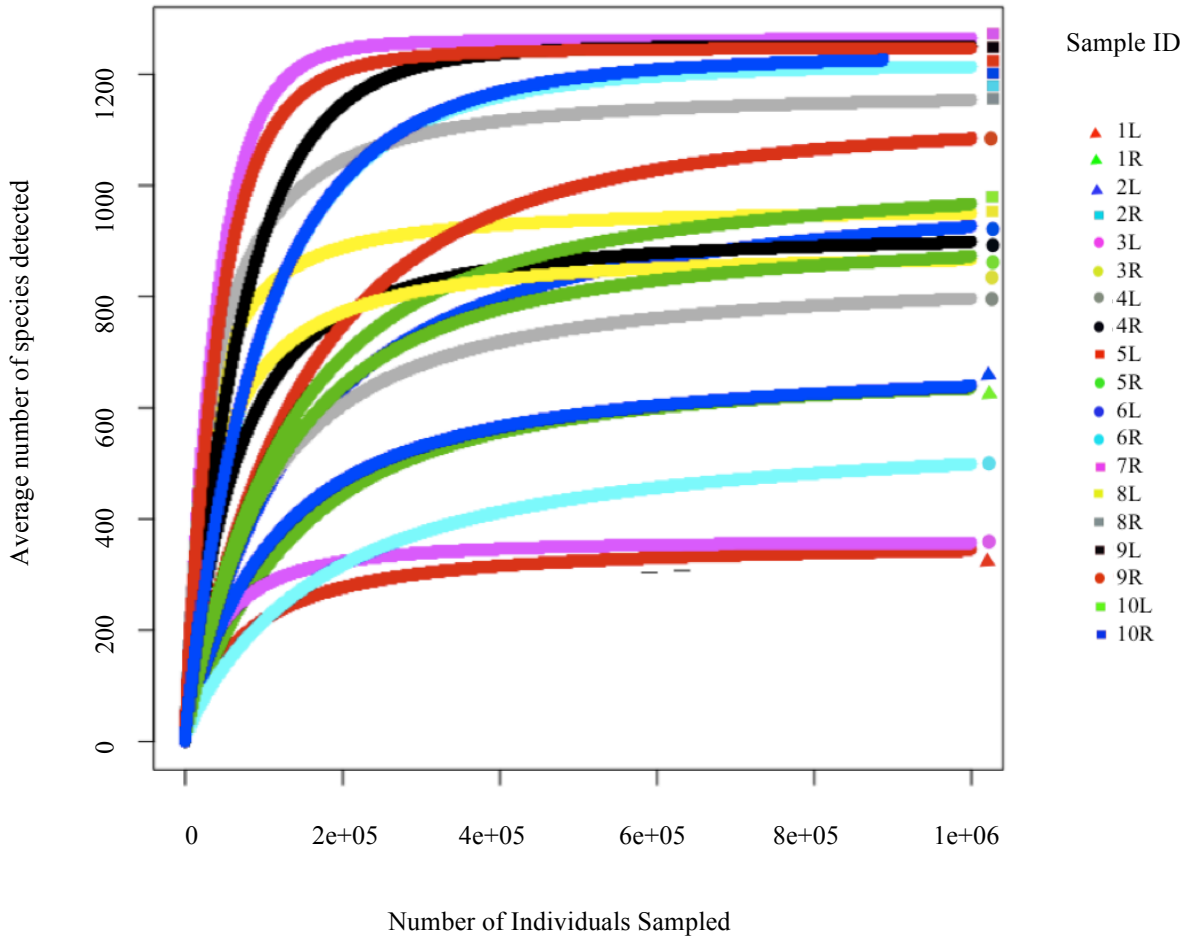


Figure 3.4. Rarefaction analysis of the 19 axillary datasets. Species richness is estimated as a function of the number of individuals (bacteria) sampled. Rarefaction analysis was calculated using the Community Ecology Package 'vegan' in R.

3.3.6 Interpersonal vs. intrapersonal microbial community variation

To measure the level of compositional similarity between all samples, Bray-Curtis dissimilarity coefficients were calculated for pair-wise combinations of all samples. A coefficient of nought indicates identical taxonomic profiles, with corresponding membership and relative abundance levels for all microbial clades, whilst a coefficient of one represents extremely dissimilar taxonomic profiles. The taxonomic composition of eleven samples (2L, 2R, 3L, 3R, 5L, 5R, 8L, 8R, 9R, 10L and 10R) were more similar to their corresponding left or right axillary sample than to any other sample (all Bray-Curtis coefficients < 0.16, Table 3.3). The remaining eight samples (1L, 1R, 4L, 4R, 6L, 6R, 7R and 9L) demonstrated the highest degree of similarity with samples not originating from the corresponding left or right axilla (all Bray-Curtis coefficients < 0.3, Table 3.3). This implies that subjects P2, P3, P5, P8 and P10 all exhibited greater interpersonal than intrapersonal variation, that is to say they were more similar to

themselves than to other participants, whilst subjects P1, P4 and P6 were more similar to other participants than to themselves. Subject P9 exhibited an unusual pattern of similarity with the right axillary profile (9R) more similar to the corresponding left axillary profile (9L) than to any other sample, however the left axillary profile (9L) exhibited a higher degree of similarity to sample 10R than the corresponding right sample (9R).

3.3.6.1 Understanding the high intrapersonal variation of subjects P1, P4 and P6

Although the taxonomic composition of both axillary samples from subject P1, (1L and 1R), were shown to be more similar to samples originating from other subjects, the Bray-Curtis coefficient of dissimilarity between the corresponding left and right axillary samples was still very low (Bray-Curtis coefficient = 0.024447727). This indicates that although the taxonomic profiles of the left and right axillae of subject P1 were more similar to other subjects, they still exhibited a high degree of similarity with each other.

Contrastingly, subjects P4 and P6 exhibited large coefficients of dissimilarity between the left and right axillae, indicating a considerable divergence between the taxonomic compositions of their corresponding axillary communities (Bray-Curtis coefficients ≥ 0.5). To understand the nature of the compositional variation between the left and right axillary communities of subjects P4 and P6, all species exhibiting statistically different relative abundances were identified using Fishers exact test with the Storey false discovery rate correction (FDR), followed by filtering out of non-biologically significant taxa by setting a ratio of proportion limit effect size of 5 ($p < 0.01$). This allowed identification of microbial clades whose relative abundance differed by a ratio of at least five between the left and right axillary communities of the same individual.

Using the described statistical tests, 71 and 79 differentially abundant species were identified between the left and right axillae of subjects P6 and P4 respectively, demonstrating a surprising similarity between the number of over and/or under represented species between both individuals. *Corynebacterium* spp. and *Staphylococcus* spp. accounted for the majority of the observed difference in relative abundances in both individuals (Fig 3.5). Different species profiles accounted for the variation between the left and right axilla of both individuals with *S. hominis* and *S. capitis* exhibiting the greatest difference in relative abundance in individuals four and six respectively (Fig 3.6). Interestingly, subject P6 exhibited contrasting malodour levels in their left and right axillae, with the left axilla (6L) emitting a low intensity of

malodour and the right axilla (6L) exhibiting a high intensity of malodour, suggesting a possible association between the compositional differences and the contrasting levels of malodour (Table 3.2). The left and right axillae of both subjects P1 and P4 were classified as emitting low levels of axillary malodour (Table 3.2).

The majority of individuals exhibited very similar taxonomic profiles in their left and right axilla and generally a higher level of interpersonal variation was observed, however unlike previous studies this work has also shown that in certain individuals the taxonomic composition of the left and right axillary microbiota can vary substantially, and a higher level of intrapersonal variation is sometimes present [22,52]. As only a small proportion of the individuals exhibited greater within-person variation than between-person variation it is possible that such individuals were not sampled by previous studies, and the fact that both studies utilised much lower-throughput technologies and different statistical analysis could also account for the deviation between results [22,52].

Table 3.3. The Bray-Curtis Coefficient of Dissimilarity displayed for each of the 19 axillary samples and the corresponding samples exhibiting the lowest dissimilarity coefficient. A low dissimilarity coefficient indicates a similar taxonomic composition. Bray-Curtis Coefficients were calculated using the QIIME community analysis software.

Sample ID	Subject ID	Sample with most Similar Taxonomic Profile	Bray-Curtis Coefficient of Dissimilarity
1L	P1	9R	0.013983099
1R	P1	9R	0.023956537
2L	P2	2R	0.051996662
2R	P2	2L	0.051996662
3L	P3	3R	0.077858008
3R	P3	3L	0.077858008
4L	P4	5R	0.175642924
4R	P4	6L	0.29783534
5L	P5	5R	0.159228722
5R	P5	5L	0.159228722
6L	P6	5L	0.191146824
6R	P6	5R	0.279293928
7R	P7	6R	0.279293928
8L	P8	8R	0.066139933
8R	P8	8L	0.066139933
9L	P9	10R	0.048649224
9R	P9	9L	0.1505736
10L	P10	10R	0.010510008
10R	P10	10L	0.010510008

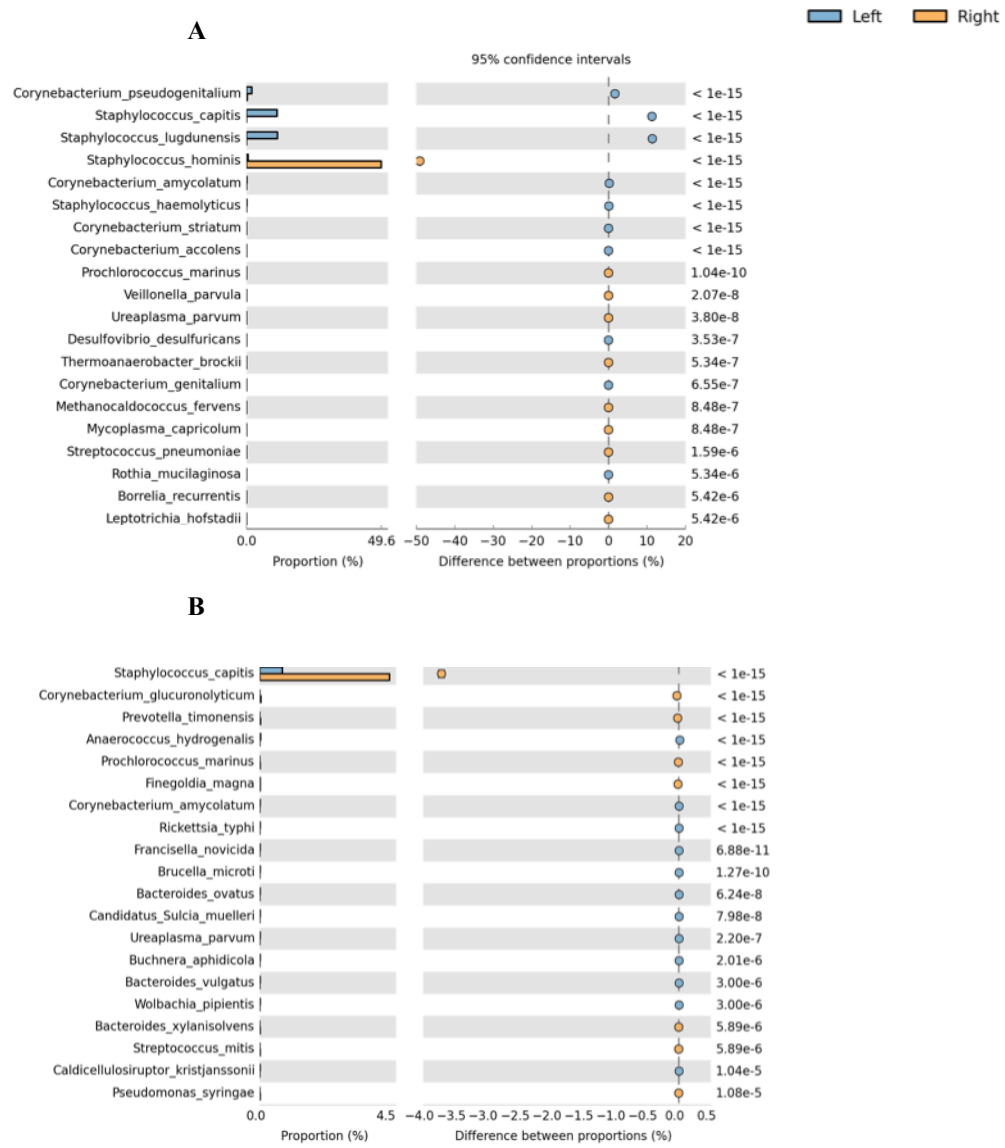


Figure 3.5. Microbial clades exhibiting differential relative abundances between the left and right axillary communities of (A) subject P4 and (B) subject P6. The 20 microbial clades exhibiting the largest differences in relative abundance are shown. Statistics and images were generated using STAMP.

3.3.7 Microbial community composition of high and low malodour axillary samples

Comparison of the microbiota from high and low malodour axillae revealed a greater average microbial diversity in higher malodour communities, with an average of 1055 ± 273 species per sample. There was no association between the dominant axillary species and the malodour intensity, as *S. epidermidis, S. hominis* and *P. acnes* were all present as the dominant species in both high and low odour samples. To understand the relationship and level of similarity between high and low malodour axillary microbial communities, all samples underwent an additional PCA analysis (Fig 3.6). Samples tended to cluster based on the identity of the dominant species within the community rather than on the level of axillary malodour (Fig 3.6). Three distinct groups were observed, each dominated by the same species and comprising microbial communities isolated from both high and low malodour environments (Fig 3.6). A certain degree of clustering was observed for all but two of the low malodour samples, indicating an overlap between their taxonomic profiles, however the high malodour samples did not cluster distinctly from the low odour group, and instead were more similar to communities with the same dominating species (Fig 3.6).

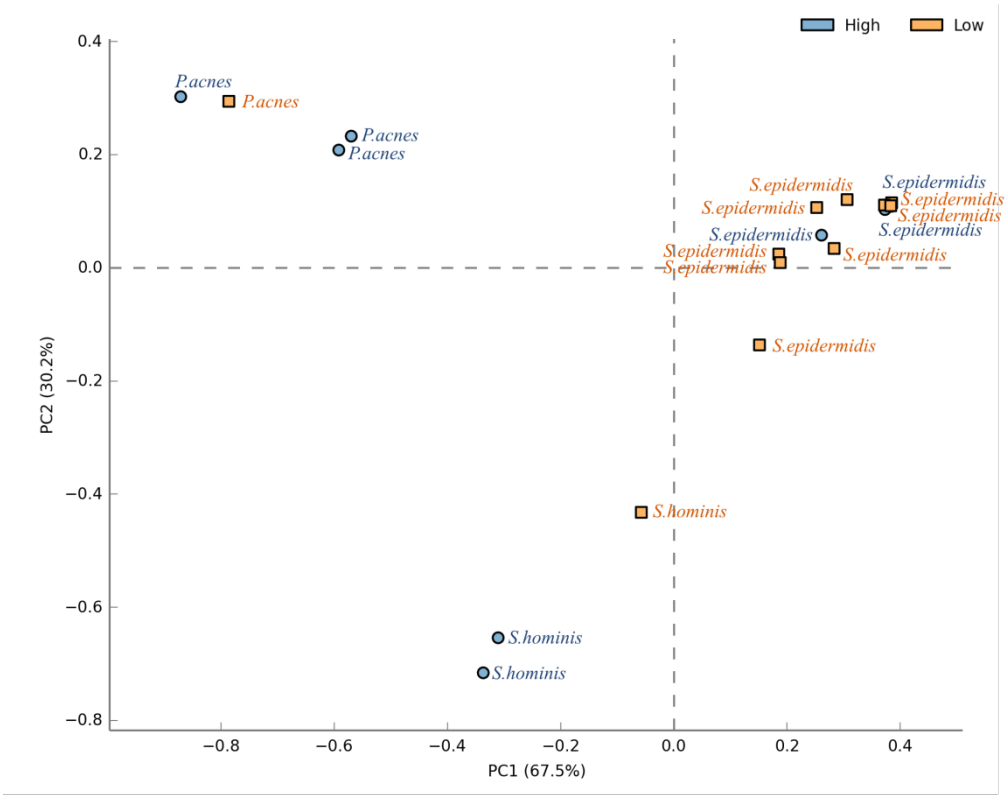


Figure 3.6. Principal component analysis of species abundance profiles from high and low malodour axillary samples isolated from 10 subjects. The most dominant species in each community is labelled next to the associated point. Each axis represents the specific principal component and the proportion of variation represented by that principal component in brackets. Principal components one and two represent the majority of the variation therefore no other principal components are shown. PCA analysis and image was generated using the STAMP statistical software.

3.3.8 Identification of enriched microbial clades within high and low malodour samples

Previous studies investigating the relationship between malodour levels and the composition of the axillary microbiota utilised culture-dependent approaches, examining the presence and abundance of certain groups of bacteria isolated from axillary samples via selective media, and associating the presence of certain taxa and bacterial numbers with levels of axillary malodour [29,268]. An abundance of certain groups of bacteria have been associated with specific types of malodour, a higher abundance of *Corynebacterium* spp. was linked to the generation of a stronger more apocrine type of malodour whilst an abundance of staphylococcal species resulted in a less pungent acid malodour [29,268]. Due to the advancement of high-throughput sequencing platforms and the ability to generate sequencing libraries from picogram amounts of DNA, this study was able to apply a completely culture-independent approach to probe the composition of the axillary microbiota and understand its association with malodour. As a

result of these techniques it was possible to compare high and low malodour microbial communities and identify biologically relevant microbial clades that were significantly associated with the production of high or low axillary malodour.

To understand which taxonomic groups were most likely to account for the difference in malodour, relative abundance profiles of all samples were analysed by a linear discriminant analysis (LDA) effect size (LEfSe) tool which uses multiple statistical tests and an LDA effect size filter to identify microbial clades significantly associated with high and low odour samples [298]. The effect size estimates the degree of responsibility associated with each microbial clade with respect to high or low axillary malodour.

LEfSe identified 59 microbial clades that were significantly enriched within high malodour microbial communities with an LDA score of two or above, indicating these groups were significantly associated with the production of high axillary malodour (Fig 3.7). Although a large number of clades were found to be associated with high malodour, only 13 were classifications to genus-level or below. Due to the considerable degree of variation observed between species with regards to their enzymatic capabilities of catalyzing odorant precursors, it is more informative to identify microbial clades associated with malodour at the lowest taxonomic level possible. Nine genera and four species were significantly enriched within high malodour samples: *Finegoldia, Pseudomonas, Kocuria, Escherichia, Streptomyces, Mycobacterium, Peptoniphilus, Burkholderia, Bifidobacterium, Finegoldia magna, Corynebacterium amycolatum, Corynebacterium kroppenstedtii* and *Kocuria rhizophila* (Fig 3.7).

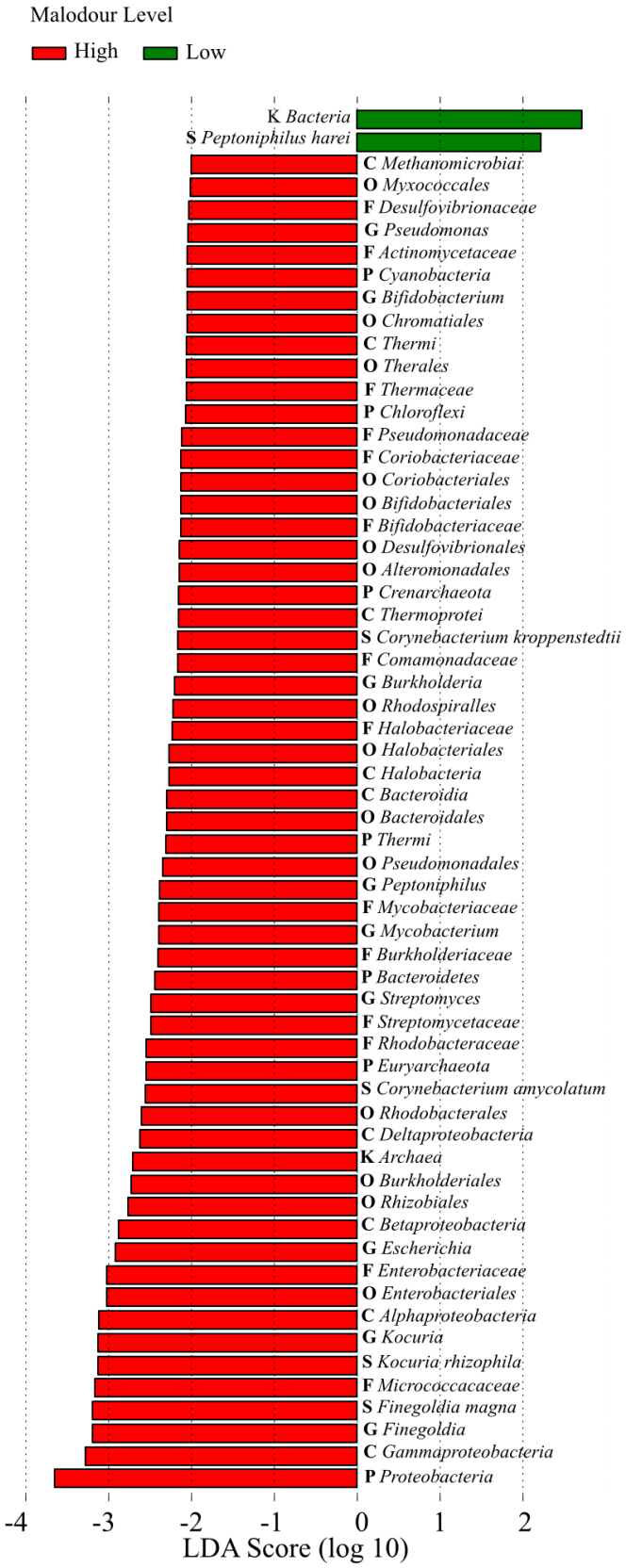


Figure 3.7. Significantly enriched microbial clades within either high or low malodour datasets exhibiting a minimum LDA score of two. Analysis and image was generated using LefSe. K, P, C, O, F, G and S refer to Kingdom, Phylum, Class, Order, Family, Order, Genus and Species.

3.3.8.1 *Finegoldia* spp.

Finegoldia spp. and *F. magna* exhibited the greatest effect size of all genera and species indicating the highest degree of association with malodorous samples. *Finegoldia* is a prevalent member of the skin microbiota and has also been isolated from the oral cavity and the gastrointestinal and urogenital tracts in significant abundances [27,38,303,304]. No previous reports have observed an association between *Finegoldia* spp. and axillary malodour, therefore it is not known whether any species possess the enzymatic capabilities to generate malodorous compounds from odourless precursors. To investigate a possible role of *F. magna* in axillary malodour generation, the whole genome sequence of *F. magna* strain BVS033A4 was examined for homologs of genes proposed to contribute towards axillary malodour *in vitro*. One of the most extensively characterised malodour formation pathways in terms of the specific microbial enzymes required is the generation of the structurally unusual volatile fatty acids (VFAs), 3M2H and HMHA, which are secreted bound to L-glutamine conjugates and subsequently released following the action of a corynebacterial Nα-acylglutamine aminoacylase (*agaA)* [72,278,279]. A potential homolog of *agaA* was identified in *F. magna,* which shared over 33% sequence identity with the *Corynebacterium* strain Ax20 *agaA* protein and encoded a putative aminohydrolase. Although there is no experimental data to validate its biochemical role in the production of an acylglutamine aminoacylase and the subsequent role in the release of malodorous VFAs, it provides a possible mechanism by which *F. magna* may contribute to axillary malodour.

3.3.8.2 *Corynebacterium* spp.

Corynebacterium spp. are frequently implicated as the major causative agents of axillary malodour and certain species possess the enzymatic capability to generate a large range of malodorous compounds from odourless precursor molecules [305]. It was not surprising therefore to find that two of the four species significantly enriched within high malodour communities were *Corynebacterium kroppenstedtii* and *Corynebacterium amycolatum* (Fig 3.7). The *Corynebacterium* genus was not significantly associated with high malodour suggesting that only the action of these specific corynebacterial species contribute towards malodour (Fig 3.7). *C. amycolatum* and *C. kroppenstedtii* have not been subject to biochemical characterisation of their enzymatic ability to generate malodorous compounds from secreted precursors, however extensive characterisation of malodour generation has been carried out in other corynebacterial species such as *C. striatum* [278].

C. kroppenstedtii is a lipophilic strain that accordingly lacks any homologs to the corynebacterial *fas* fatty acid synthase gene, explaining an *in vitro* dependence on exogenous

fatty acids [306]. *C. amycolatum* was previously characterised as non-lipophilic and encodes a fatty acid synthase for the biosynthesis of lipids [307]. Lipophilic and lipid-catabolising strains have previously been associated with high levels of axillary malodour, and the partial degradation of long-chain cutaneous fatty acids (LCFAs) via the β-oxidation pathway is thought to be a major mechanism of malodour generation. [268,275]. Since homologs of all genes involved in the β-oxidation pathway were previously bioinformatically characterised in *C. kroppenstedtii,* it is possible that this mechanism is utilised by *C. kroppenstedtii* to partially degrade cutaneous fatty acids to VFAs and may play a major role in generating malodour [275].

Although no studies have associated non-lipophilic corynebacteria with axillary malodour, as some strains possess the enzymatic capability to degrade fatty acids it is possible that this group of bacteria play a role in VFA generation. To understand if the non-lipophilic *C. amycolatum* may be involved in LCFA degradation, the whole genome sequence of *C. amycolatum* strain SK46 was searched for homologs of the four genes required for β-oxidation: *fasA, fasB, fasE* and *fasH*. Amino acid sequences sharing at least 55% sequence identity were found for all four β-oxidation enzymes in this strain revealing the presence of homologous genes. This indicates that although *C. amycolatum* is non-lipophilic and does not require exogenous fatty acids for growth, it may possess the enzymatic ability to degrade fatty acids via β-oxidation. The enrichment of *C. amycolatum* within high malodour samples indicates this mechanism may be employed as a route towards generating malodorous VFAs. The lack of axillary isolated corynebacteria with the ability to degrade fatty acids had recently been quoted as a reason to question the extent to which partial LCFA degradation contributes to axillary malodour, however the discovery that two *Corynebacterium* spp. significantly associated with high axillary malodour both contain the required genes for β-oxidation indicates that this mechanism may be more accountable for axillary malodour than previously thought.

Interestingly, another similarity shared by *C. amycolatum* and *C. kroppenstedtii* is the lack of mycolic acid in the cell envelope, which is a fundamental physiological component of all other human isolated corynebacterial species apart from *C. atypicum* [306,308-311]. Accordingly, both species lack a *fadD1* gene, which is involved in fatty acid biogenesis via the synthesis of mycolic acid.

Volatile sulphur compounds (VSCs) such as 3-methyl-3-sulfanylhexan-1-ol (3M3SH) contribute to axillary malodour due to their low olfactory threshold [75,76,88]. They are secreted bound to Gly-Cys-(S)-conjugates and subsequently released by the enzymatic action of corynebacterial and/or staphylococcal species of the axillary microbiota [287]. The precise mechanism by VSCs are cleaved from their precursors has not been fully characterised in

staphylococci, however the combined action of two corynebacterial enzymes, a C-S β-lyase (*Aec*/*MetC*) and a metal dependent dipeptidase (*TdpA*), are able to generate 3M3SH from Gly-Cys-(S)-conjugates [287].

Investigation of the whole genome sequences of *C. amycolatum* and *C. kroppenstedtii* revealed homologs of the *C. jeikeium* C-S β-lyase amino acid sequence in both species, with percentage identities over 40%, indicating that both genomes encode enzymes with similar functions. Correspondingly, homologs of the corynebacterial dipeptidase required alongside C-S β-lyase for release of malodorous thiols from Gly-Cys-(S)-conjugates were also identified in both corynebacterial genomes at similar sequence identity levels. Although the enzymatic ability of *C. amycolatum* and *C. kroppenstedtii* to release malodorous thiols from Gly-Cys-(S)-conjugates has not been subject to experimental validation, the presence of genes encoding for both required enzymes indicates a similar mechanism may occur in both species. Due to their significant enrichment within high malodour samples it is possible that this mechanism is a major cause of axillary malodour production.

The final mechanism by which corynebacterial species contribute towards axillary malodour production is by the release of structurally unusual medium and short chain fatty acids such as 3M2H and HMHA from glutamine bound precursors[72,88,279]. This mechanism has been previously been fully characterised in *Corynebacterium* strain Ax20, in which the enzymatic action of the an aminoacylase (*AgaA*) was shown to release 3M2H from its glutamine bound conjugate [278]. Homologous proteins sharing over 36% sequence identity with the original *C. striatum* gene were found in the genomes of both *C. amycolatum* and *C. kroppenstedtii* during this study, indicating the possibility that both species encode an enzyme with similar glutamine cleaving activities. The prevalence of structurally unusual medium and short chain fatty acids in malodorous sweat indicates that the production of 3M2H and HMHA is a primary mechanism of malodour generation, and as *C. amycolatum* and *C. kroppenstedtii* are both significantly associated with high malodour samples and both contain the hypothetical enzymatic ability to generate 3M2H and HMHA, it is likely that this mechanism is utilised by these two species to generate malodorous compounds within the axillary vault.

Both of the enriched corynebacteria are poorly characterised with respect to their enzymatic ability to generate malodorous compounds. The extensive characterisation of other corynebacterial species such as *C. jeikeium* and *C. amycolatum* has led to identification of the genes important in malodour generation. This information has subsequently allowed predictions to be made regarding the functional potential of *C. striatum* and *C. kroppenstedtii* with respect to malodour generation pathways.

3.3.8.3 *Kocuria* spp.

The fourth species significantly enriched within high malodour samples was the Gram-positive actinomycete *Kocuria rhizophila. T*he genus *Kocuria* was also significantly associated with high malodour samples, therefore it is possible that this group of bacteria make a considerable contribution to axillary malodour. *K. rhizophila* is a strict aerobic coccoid that was first described in 1999 and since then has been infrequently reported in the literature. It is normally isolated from soil, although recently it was identified as the etiological agent of a human infection as a result of a persistent colonisation of an indwelling medical device [312-314]. Although there are no other reports of a human originating *K. rhizophila* isolate, other members of the *Kocuria* genus are considered normal inhabitants of the human microflora, and were recently shown to colonise deeper layers of the stratum corneum [21,55]. This is the first report of the presence of *K. rhizophila* as a member of the axillary microbiota, and the first indication that this species or genus may play a role in the generation of malodorous compounds in the axilla. It is possible that the novel approach utilised by this study has revealed a previously unknown inhabitant of the axillary microbiota.

3.3.8.4 A diverse range of genera are implicated in axillary malodour generation

Alongside the four species-level microbial clades, the nine following genera were also significantly enriched within high axillary malodour samples: *Finegoldia, Pseudomonas, Kocuria, Escherichia, Streptomyces, Mycobacterium, Peptoniphilus, Burkholderia* and *Bifidobacterium* (Fig 3.7). Many of these genera are not considered normal members of the human skin microbiota, and even fewer have been identified or isolated from the axillary vault. *Escherichia* and *Bifidobacterium* are both commonly found in the gastrointestinal tract where they exist as commensal inhabitants of the gut microbiota, however both have occasionally been identified as transient members of the skin microbiota [48,55,315]. Neither genera has been previously associated with the axillary microbiota or the generation of malodorous compounds, however as *E. coli* is able to catabolise fatty acids via β-oxidation and *Bifidobacterium* spp. are able to generate multiple volatile fatty acids including butyric acid and acetic acid, it is possible that both may contribute towards axillary malodour [316,317]. Due to their infrequent isolation from the axillary microbiota it is more likely that both genera are not generally involved in malodour generation and they are either present as transient members of the axillary microbiota or as a result of post-sampling contamination. *Burkholderia* and *Mycobacterium* were also significantly enriched within high malodour axillary samples, however neither are common

members of the skin or axillary microbiome. Both are frequently associated with human infections including cystic fibrosis, skin infections and soft tissue infections [116]. Although not members of the skin microbiome there have been infrequent reports of *Mycobacterium* isolates originating from the axillary vault, indicating they may be less abundant permanent members of the axillary microbiota [318].

Kocuria, Pseudomonas, Finegoldia and *Peptoniphilus* are frequently isolated from human skin and are considered prevalent members of the skin microbiome [27,55,82,319]. Both *Peptoniphilus* and *Finegoldia* are able to produce short chain fatty acids, and as the generation of VFAs is a major source of malodour it is possible that the enzymatic activity of *Peptoniphilus* and *Finegoldia* species upon odourless precursors within axillary secretions may contribute to malodour within the axilla [320,321]. Although there is no experimental evidence to associate *Kocuria* and *Pseudomonas* species with malodour generation, their significant enrichment within high malodour samples indicates a possible contribution.

None of the dominant species identified in any of the samples were significantly associated with either high or low malodour, indicating that the presence of *S. epidermidis, S. hominis* or *P. acnes* as an abundant organism does not indicate or predict a specific level of axillary malodour. This also suggests that the most abundant organism in the community may not assume the greatest responsibility for the generation of malodour. Although *Corynebacterium* spp. were not present at a high relative abundance in any samples, two corynebacterial species were identified as significantly enriched within high malodour samples, indicating they may contribute towards axillary malodour generation at very low abundances levels.

3.3.8.5 *P. harei* is enriched within low axillary malodour samples

Only two microbial clades were significantly associated with low axillary malodour. As LEfSe identifies microbial clades at all taxonomic levels some of the enriched groups at higher taxonomic levels were less informative, for example one of the enriched clades within low malodour samples was the kingdom *Bacteria,* whilst an enriched group within high odour samples was *Archaea* (Fig 3.7). Ignoring the kingdom level classification, the only enriched group within the low odour samples was the Gram-positive anaerobic coccus (GPAC), *Peptoniphilus harei*. *P. harei* is present as one of the top 20 most abundant species in 58% of the axillary samples and has previously been reported as a common skin isolate and a prevalent member of the axillary microbiota [50]. Its enrichment within low malodour communities may be

a consequence of the absence of one or more microbial clades which usually inhibit its dominance in high malodour communities.

3.3.9 Functional analysis of high and low malodour microbial communities

3.3.9.1 IDBA-UD metagenomic assembly

Utilising a whole-genome approach to analyse the axillary microbiota not only generated an more un-biased representation of the taxonomic content of the axilla in comparison to using a 16S rRNA approach, but also provided access to the extensive gene content of the microbial community, allowing inferences to be made regarding their possible functional roles. To generate the most accurate functional profile possible, all filtered reads were assembled into longer contiguous sequences and subject to functional annotation. In chapter two the accuracy of two metagenomic assembly tools, MetaVelvet and IDBA-UD, were compared using an *in vitro* synthetic community of known composition [182,183]. Based on subsequent annotation of the resulting contigs it was found that contigs assembled by MetaVelvet provided the most comprehensive representation of the functional potential of the microbial community, therefore all axillary datasets generated in this chapter were assembled using this tool. The resulting assemblies varied considerably in size with N50 lengths ranging from 700 bp to 9,000 bp, with an average N50 of 2,200 bp (Table 3.4). The percentage of reads incorporated into the contigs was relatively high for every dataset, with no less than 60% of filtered reads incorporated into the final assembly of each dataset.

To understand the functional profile of high and low malodour microbial communities all assembled datasets were subsequently subject to structural and functional annotation via the IMG/M-ER metagenome annotation pipeline. An average of 30,888 ± 12,784 protein coding genes were identified per dataset, of which 63% ± 7% were classified to a Cluster of Orthologous proteins Group (COG) (Table 3.4). The unclassified genes within each dataset are likely to represent novel genes with no known homologs within the COG database. The number of genes classified to each COG was calculated for each dataset and normalised based on the total number of COG annotated genes to allow direct comparison between samples.

Table 3.4. Assembly and annotation statistics for 19 axillary datasets isolated from the left and right axilla of ten subjects. Filtered datasets were assembled using the de-Bruijn graph-based assembly tool MetaVelvet, and assembled contigs were submitted to IMG/M for functional annotation. The percentage of assembled reads is expressed in relation to the total number of filtered reads, and the percentage of genes generating COG alignments is relative to the total number of protein coding genes.

	MetaVelvet Assembly			IMG/M-ER Functional Annotation	
Sample ID	**N50 Length (bp)**	**Total Contig Length (bp)**	**Assembled Reads (%)**	**Protein Coding Genes**	**Genes Generating COG Alignments (%)**
1L	9,045	3,018,738	85.79	4,416	61.44
1R	1,080	9,430,766	89.20	18,151	64.92
2L	2,077	12,405,813	91.01	23,803	63.84
2R	1,604	15,395,649	89.33	27,780	66.62
3L	1,180	9,718,084	66.29	19,968	59.57
3R	911	10,302,489	62.26	24,652	57.15
4L	2,351	22,540,706	87.41	40,543	67.20
4R	1,592	11,596,924	91.16	22,322	65.78
5L	875	19,507,998	84.90	40,668	65.60
5R	2,651	22,556,938	86.48	39,840	65.38
6L	1,865	20,404,504	86.99	36,422	44.55
6R	1,573	25,321,973	85.20	46,473	46.25
7R	1,750	25,386,651	81.64	48,523	66.67
8L	694	16,673,034	71.74	37,080	65.57
8R	1,189	23,371,605	75.56	45,173	66.59
9L	1,831	17,871,643	89.94	31,000	68.28
9R	915	22,491,674	78.37	47,245	68.45
10L	1,865	9,132,778	99.09	17,870	65.32
10R	6,462	9,381,783	94.05	14,937	67.54

3.3.9.2 Enrichment of specific genes within high and low malodour datasets

To understand which COG groups were significantly enriched within high and low odour datasets, relative abundance distributions were analysed using the previously described LDA effect size tool, LEfSe [298]. Surprisingly, the number of differentially abundant COGs was relatively low in both high and low odour groups, with 27 COGs enriched within low odour datasets and only 16 COGs enriched within high odour datasets (Fig 3.8). Of the 16 COG groups enriched within high malodour samples, three were associated with nucleotide transport and metabolism (COG0044 COG0046 and COG0402), three with amino acid transport and metabolism (COG1104, COG2986 and COG4166), two with translation, ribosomal structure and biogenesis (COG0060 and COG1185), two with cell envelope biogenesis, outer membrane (COG2247 and COG0773), two with posttranslational modification, protein turnover and chaperones (COG0606 and COG0520) and one with each of the following: defence mechanisms (COG1132), DNA replication, recombination, and repair (COG0587) and signal transduction mechanisms (COG1966) (Fig 3.8).

Investigation of the predicted functions and associated enzymatic pathways of all 16 enriched COGs revealed two groups with possible roles in malodorous compound generation. Groups COG1104 and COG0520 were classified as '*cysteine sulfinate desulfinase/cysteine desulfurase and related enzymes'* and '*selenocysteine lyase'* respectively. Following investigation of all the genes generating significant alignments against each COG it was found the majority of aligned genes within the dataset were annotated as cysteine desulfurases, with COG1104 aligned genes annotated as *IscS* and COG0520 aligned genes annotated as *SufS*. Both genes encode pyroxidal-5'-phosphate (PLP) containing enzymes that are involved in the two major prokaryotic iron-sulphur (Fe-S) cluster biogenesis pathways, Suf and Isc [322]. Fe-S clusters are one of the most predominant prosthetic groups in biology and are involved in a diverse range of functions [322]. The Suf (sulphur formation) system comprises a six gene operon *sufABCDSE,* whilst the Isc (iron sulphur containing) operon consists of four genes *IscSRUA* [323,324]. Within their respective pathways SufS and IscS are both characterised as cysteine desulfurases which are responsible for the conversion of L-cysteine to L-alanine and free sulphur which subsequently forms a persulfide intermediate required for the next stage of Fe-S cluster assembly [323,325,326]. Due to their significant enrichment within high malodour samples it is possible that the action of these bacterial enzymes on L-cysteine from axillary secretions may contribute towards axillary malodour. Although this specific enzymatic capability has not been demonstrated in axillary isolated bacteria, a recent study characterising all corynebacterial aminotransferases identified a protein with homology to SufS in *C. glutamicum* and found that it had the enzymatic ability to

generate malodorous sulphur derivatives from L-cysteine substrates [327]. Also, homologs of both *sufS* and *IscS* were identified in the genome sequences of *C. amycolatum* and *C. kroppenstedtii,* the two corynebacterial species previously implicated as major contributors of axillary malodour, and although not experimentally validated it is likely that both species possess the enzymatic ability to generate malodorous sulphur derivates via this mechanism. Due to the significant enrichment of these genes and the two *Corynebacterium* spp. *C. amycolatum* and *C. kroppenstedtii* in high malodour samples, it is possible that both species utilise this mechanism to generate malodorous sulphur derivatives. Although the role of these enzymes has not been characterised in relation to axillary malodour production, the contribution of volatile sulphur compounds (VSCs) to axillary malodour has been extensively documented, and as VSCs have a very low olfactory threshold only a small concentration is required to produce a significant malodour. The most documented mechanism resulting in sulphur derived compounds is the cleavage of sulfanylalkanols 3M2H and HMHA from their Gly-Cys-(S)-conjugate precursors via the action of a corynebacterial C-S lyase (AecD) and metal-dependent dipeptidase (TdpA). It is possible that these two enzymes represent an additional VSC production pathway that is utilised by axillary bacteria, and contributes to the well-characterised array of malodorous sulphur derivatives present in axillary sweat.

Addition enriched functions within high malodour samples included multiple ABC-type transport systems, genes involved in purine and pyrimidine metabolism, ligases involved with peptidoglycan biosynthesis and carbon starvation response genes. Further characterisation into malodour generation is required to understand if or how these genes may influence the generation of malodorous compounds.

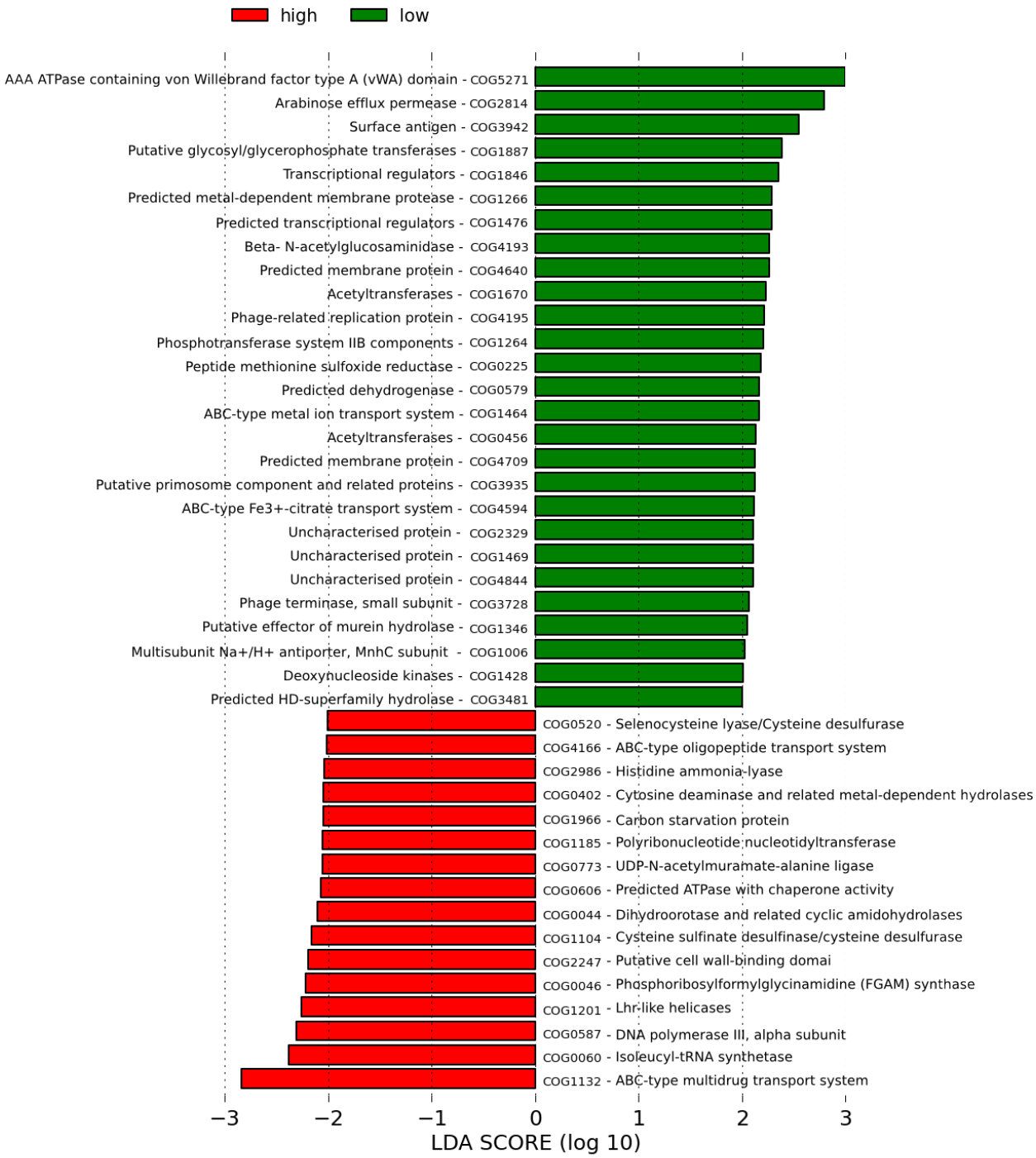


Figure 3.8. COG groups significantly enriched within either high or low malodour datasets with an LDA cut-off of two. Differentially abundant COG groups were identified using the LDA effect size tool LEfSe.

3.4 Conclusion

Using a large group of subjects, this work generated a comprehensive description of the taxonomic and functional composition of the axillary microbiota. Based on whole-genome metagenomic data, the taxonomic profile of the axilla was revealed to be very distinct in structure, comprising one dominant species belonging to either the *Staphylococcus* or *Propionibacterium* genera and a high number of less abundant microbial clades. Although previous studies have also characterised an abundance of staphylococcal species within the axilla, the predominance of propionibacteria over corynebacterial species has never been reported [268,271]. In fact, this study revealed that *Corynebacterium* were not dominant in any samples, and no corynebacterial species accounted for more than 6% of the total community abundance. A high abundance of corynebacteria has been a long accepted characteristic of the axillary microbiome due to its frequent isolation from the axillary vault, and the characterised ability of a number of corynebacterial species to generate malodourous compounds [52,73,78,268,271]. The evidence presented by this study suggests that *Corynebacterium* is not a predominant genus within the axilla, and that *Propionibacterium* spp. are actually more abundant. The novel conclusions reached by this study may be a result of the unbiased approach utilised to profile the axillary community, as all previous studies employed 16S rRNA-based techniques, which have inherent associated biases [22,268,271]. The impact of primer selection in particular has been shown to considerably impact the resulting species diversity estimations [302,328,329]. However, as highlighted in chapter two, in which an *in vitro* simulated microbial community was subject to whole genome shotgun sequencing, a metagenomic approach does not predict a completely accurate representation of the taxonomic content of a microbial community, and may lead to the over/under representation of certain microbial clades. It is also possible that there are unknown biases associated with the use of the novel transposition-based library preparation technique or the extremely low level of input DNA. As this is the first application of metagenomic sequencing to the axillary microbiota, further studies of a similar nature and on a larger scale are required to more accurately define the taxonomic composition of the axillary microbiota.

The generation of axillary malodour is associated with abundance of either corynebacterial, staphylococcal or less commonly, propionibacteria species within the axillary microbial community, which are responsible for the biotransformation of secreted odourless precursors to malodorous compounds. This study found no association between any of the dominant taxonomic groups and high malodour, and instead revealed the significant association of the four low-abundance species *F. magna, C. amycolatum, C. kroppenstedtii* and *K. rhizophila*, with the generation of high axillary malodour. Upon investigation of their individual contributions to community abundance, it was found that none of the enriched species

accounted for more than 0.75% of the entire microbial community of any sample, suggesting that certain rarer species may exhibit a more significant contribution to axillary malodour than abundant species. Less abundant species are often implicated in ecologically important roles within microbial communities. For example, a species comprising only 0.006% of the total abundance of a peatland-associated microbial community was found to play a major role in sulphate reduction [330]. In addition low abundance community members are often key components of complex inter-species interactions involving more abundant species.

It was of significance to note that two of the enriched species within high malodour communities belonged to the *Corynebacterium* genus, confirming the suspected role of corynebacterial species in malodour generation. Numerous studies have characterised a variety of mechanisms by which certain corynebacterial species can generate malodourous compounds [73,74,78,79,269]. Although the biochemical capabilities of the two enriched species *C. amycolatum, C. kroppenstedtii,* are not currently understood in relation to the generation of malodourous compounds, investigation of the genome sequences of both species revealed the genetic equipment required for numerous enzymes involved in malodour generation, including a corynebacterial aminoacylase, C-S β-lyase, metal dependent dipeptidase and β-oxidation pathway, indicating their possible contributions to axillary malodour. This study presents a novel hypothesis regarding the abundance of *Corynebacterium* spp. within the axilla and the relationship between relative abundance and the generation of malodour, suggesting that although no corynebacterial species dominate the community, they still play a major contributory role towards malodour generation.

Until now, our understanding of the specific microbial genes involved in malodour generation has been based upon culture-dependent studies involving single groups of isolates [72,73,78-80,88,278]. Although these studies have dramatically increased our understanding of the microbial pathways required for the release of malodourous compounds, they do not reflect the functional activity of the axillary microbiota *in vivo*. By utilising a whole-genome approach, this study identified two microbial cysteine desulfurases, *sufS* and *iscS*, which were significantly enriched within high malodour samples. As both enzymes lead to the generation of sulphur-derived compounds, they may represent novel mechanisms by which the axillary microbiota are able generate malodour. Future work will be required to determine the transcriptional status of *sufS* and *iscS* within high malodour communities to understand if they actively contribute to axillary malodour. Metatranscriptomics is a complementary profiling technique that reveals the active cellular processes within a community by direct extraction of community RNA. This technique has been successfully used to profile many soil and ocean associated microbial communities, however its application to the skin microbiota has been limited due to the low sample biomass

[331-334]. It is hoped techniques utilised in this study to overcome low DNA yields may also be applicable to generate cDNA libraries from low level microbial RNA samples, paving the way for future studies.

This chapter presented the first unbiased view of the taxonomic and functional composition of the axillary microbiome by utilising a whole-genome metagenomics approach. It is hoped that this work will provide a starting point for the application of future 'omics' techniques such as metatranscriptomics, metaproteomics and metabolomics, in order to more comprehensively understand the functional roles and interactions of the axillary residents.

CHAPTER 4

Pan-genome analysis of the nosocomial pathogens and common skin isolates *S. epidermidis* and *S. aureus*

4.1 Introduction

Staphylococcal species are the most abundant members of the skin microbiome, and have been isolated from the majority of skin-associated sites [21,335,336]. The most frequently encountered species on the skin is *S. epidermidis*, which exhibits the highest abundance levels in moist skin sites such as the nare (nose), axilla, antecubital fossa (inner-elbow) and the umbilicus (navel) [52]. Although it exists as a beneficial commensal on human skin, its ubiquitous nature, ability to form biofilms on inert surfaces, and retention of antibiotic resistance genes has led to it becoming one of the leading causes of medical-device associated infections [84]. *S. aureus* is another important host-associated staphylococcal species, which is less frequently isolated from the skin in comparison to *S. epidermidis*, and resides primarily in the mucosal membranes of the anterior nare [337]. Unlike *S. epidermidis*, which is thought to colonise every person on the planet, *S. aureus* residency is limited to approximately 1/3rd of the population [85]. The increased arsenal of both secreted and cell surface associated virulence factors encoded by *S. aureus* allow it to exhibit a much more pathogenistic lifestyle than *S. epidermidis*, and the widespread acquisition of antibiotic resistance genes in hospital settings has led to it emerging as one of the most common etiological agents of nosocomial infections [338-341].

4.1.1 Background of *S. epidermidis* colonisation and infection

The presence of *S. epidermidis* on the skin confers a myriad of beneficial effects upon the host such as protection against the colonisation of pathogenic species and stimulation of the host immune response [61]. It is thought that the ubiquitous nature of *S. epidermidis* has led to it emerging as the primary causative agent of indwelling medical device associated infections, and a recent study determined that out of every 1,000 peripheral or central intravenous catheter insertions in the USA, 4-5 resulted in a bloodstream infection with 22% diagnosed as *S. epidermidis* [342]. As *S. epidermidis* lacks the arsenal of tissue damaging exoenzymes possessed by its more virulent relative *S. aureus,* infections are usually confined to compromised patients, such as those undergoing immuno-therapy, AIDS patients, drug users and premature new-born babies [343]. Patients who are neutropenic possibly due to chemotherapy or radiotherapy and patients with head, neck or breast carcinoma also exhibit increased susceptibility to *S. epidermidis* infections, which usually present as cellulitis with subsequent sepsis[344]. Device-associated contamination is the most common cause of *S. epidermidis* infections commonly

leading to bacteraemia, however in rarer cases it can lead to more serious conditions such as prosthetic valve endocarditis, of which it is the second most common causative agent [345]. Antibiotic resistance is becoming increasingly common amongst *S. epidermidis* nosocomial isolates, and a surveillance for methicillin-resistant isolates (MRSE) within hospitals worldwide, found that 75-90% of isolates possessed this characteristic [346]. The reduced affinity for methicillin is due to the presence of the *mecA* gene which encodes the penicillin binding protein PBP2a [347].

The success of *S. epidermidis* as a pathogen is partly attributed to its ability to form biofilms on synthetic surfaces, which provides a physical barrier against certain antibiotics and the efforts of the host immune system, allowing development of the biofilm and subsequent detachment of cells leading to bloodstream infections [348,349]. Biofilms induce an altered cell state causing a reduction in the number of active cellular processes occurring within the microbial cells. This dormant state inhibits the activities of certain antibiotics that target active cellular processes such as cellular division and DNA replication [84,343,350]. It has also been shown that biofilms severely impede the actions of the host immune response by inhibiting the actions of phagocytes and preventing the activation of the complement system [351].

Biofilm formation is divided into two stages, the initial attachment of cells to the synthetic surface followed by bacterial accumulation and the development of a mature biofilm. Primary attachment to a synthetic surface is a multi-factorial process, and therefore the roles of all proteins involved are not fully understood. Accumulation-associated protein (Aap) has been linked with biofilm formation due to the inhibitory effect exhibited by anti-aap antibodies upon *S. epidermidis* biofilm formation, however its exact role has not been fully elucidated [352]. One component of biofilm formation for which the role and structure has been extensively characterized is polysaccharide intercellular adhesin (PIA). PIA was shown to be essential for intercellular adhesion during the primary attachment phase of biofilm formation [353] PIA is synthesised by *ica*ADBC locus encoded proteins [354]. It has been shown that the Ica locus is commonly disrupted by insertion sequences creating PIA-negative *S. epidermidis* strains. In this case these strains are often still able to produce biofilms using a truncated form of Aap [352]. Other important proteins during *S. epidermidis* biofilm formation include the surface associated proteinacious autolysin AltE, and the extra-cellular matrix binding protein Embp, which has been shown to bind fibronectin in certain *S. epidermidis* strains [355] . Extracellular DNA (eDNA) also aids biofilm formation by facilitating bacterial accumulation onto the device surface [356]. Two-component signal transduction systems (TCS), which consist of a membrane associated histidine kinase and a cytoplasmic response regulator, have also been linked to biofilm formation and virulence [356]. Recently, the first *S. epidermidis* pathogenicity island was

identified and shown to express staphylococcal enterotoxin C3 (SEC3) and staphylococcal enterotoxin-like toxin L (SElL) [357]. There has also been evidence that some *Staphylococcal* toxins important in *S. aureus* infections such as toxin A (*sea),* B (*seb*), C (*sec-1*), D(*sed*) and TSST-1 may also be present in a limited number of coagulase-negative staphylococci*,* suggesting their possible presence in *S. epidermidis*, showing a need to further investigate the presence of *S. aureus* virulence genes in *S. epidermidis* [358].

Until recently, only two *S. epidermidis* genome sequences were publically available, (ATCC 12228 and RP62A) [343,359]. In the previous 12 months, 60 *S. epidermidis* draft whole genome sequences have been generated and deposited in the public databases. This drastic increase in *S. epidermidis* genomic data is due in part to the progress of the NIH funded Human Microbiome Project (HMP) in isolating and sequencing human-associated *S. epidermidis* strains [360]. As they are serving as reference genomes for future human microbiome analyses they have not undergone genomic analysis beyond assembly and basic functional annotation. A recent comparative study by Conlan *et al.,* 2012 is also responsible for the considerable increase in the number of *S. epidermidis* genome sequences, releasing the draft genome sequences of 30 *S. epidermidis* strains into the public databases [361].

4.1.2 S. aureus

S. aureus is the second most frequent causative agent of hospital-acquired (HA) blood infections and is also emerging as a significant cause of community-acquired infections (CA). This opportunistic pathogen causes a much wider and more acute range of infections in comparison to the more chronic infections caused by *S. epidermidis,* such as wound infections, pyogenic infections, bacteraemia, endocarditis, osteomyolitis, pneumonia, food poisoning, UTI infections and toxic shock syndrome (TSS)[362]. The increased virulence of *S. aureus* is due to the presence of large array of toxin encoding genes, including enterotoxins, exotoxins, leukocidins and leukotoxins, that are all encoded by genes present on multiple genomic islands known as pathogenicity islands[359]. Similarly to *S. epidermidis,* a major hindrance to the treatment of *S. aureus* infections is the extensive antibiotic resistance capability. Antibiotic resistance was first seen in *S. aureus* in the 1940's shortly after the general use of penicillin first began [363]. This was swiftly followed by the emergence of methicillin resistant strains a year after the antibiotic was introduced, and currently there are many multi-drug resistant forms of *S. aureus* such as methicillin-resistant *S. aureus* (MRSA) and vancomycin-resistant *S. aureus* (VRSA), which collectively are resistant to the majority of antibiotic classes including penicillin, chloramphenicol, macrolides, tetracycline and aminoglycosides [344-346,364-366].

Due to their frequent co-colonisation of the same human-associated niches, *S. aureus* and *S. epidermidis* strains exhibit numerous inter-species interactions, ranging from the specific inhibition of *S. aureus* colonisation by *S. epidermidis,* to horizontal gene transfer of mobile genetic elements between the two species [59]. A genomic comparison of individual *S. epidermidis* and *S. aureus* isolates revealed that a set of 1,681 putative genes are conserved and share synteny across both genomes, while the main source of variation between the two species exists in genomic islands is non-syntenic parts of the genome [359,367].

4.1.3 Pan-genome analysis

Previously, genomic characterisation of a species involved generating the whole genome sequence of one strain, which would subsequently be accepted as a reference for that species. However it is now apparent that sequencing one or two genome sequences does not provide a comprehensive description of that species, and often fails to represent the often vast intra-species diversity and variation. Due to the advancements in high-throughput sequencing technologies in the last ten years, there has been a recent surge in the number of publically available bacterial genome sequences. This increased availability of sequence data has allowed a novel type of analysis known as 'pan-genomics' to emerge, which utilises multiple strains to understand the entire gene repertoire of a species. This type of analysis was first conveyed by Tettelin *et al.* in 2005, when they described the pan-genome of a species as comprising a core set of genes present in every characterised strain known as the 'core' genome, and a set of variably present genes and strain-specific genes known as the 'accessory', 'variable' or 'dispensable' genome [139]. Mathematical modelling and extrapolation of the pan-genome can predict whether the species has an 'open' pan-genome which increases in size upon the inclusion of additional strains due to the incorporation of novel genes, or whether the pan-genome is 'closed' and does not increase in size as the full genetic repertoire of the species has already been described [139,141]. Species classified with an open pan-genome often have increased intra-species variability in comparison to those species with a closed pan-genome, and possess a larger number of accessory genes [368,369].

Since core genes are present in every identified strain of a species, the majority are unsurprisingly responsible for cellular processes required for basic survival and phenotypic characteristics of a species such as those involved in transcription, translation, replication and essential metabolic processes [370]. Core genes can account for between 30-90% of the total protein coding gene of genome depending on particular species [138-140,368,369,371]. An increased number of core genes indicates a higher degree of intraspecies conservation and little variation between strains. As the core gene group does not usually account for the majority of the

genome, to fully understand the overall intricacy of a bacterial species it is necessary to understand the genome content of multiple strains.

The accessory or dispensable genome is divided into two categories of genes, those present in two or more but not all strains, and those present in one strain only, described as singletons or strain-specific genes. Accessory genes are not conserved throughout the species, and are therefore usually responsible for functions that are not essential for survival but may enhance the ability of a species to thrive in certain environments. Functions relating to pathogenicity, resistance to antibiotics, adaptation to specific niches and motility are often encoded by the accessory genome [138-140,372]. The accessory genome is usually responsible for the majority of intra-species variation, however a small number of species displaying an alternative variation structure which exhibit the majority of variation as core-gene small nucleotide polymorphisms (SNPs) and indel regions have been identified [373]. The size of the accessory genome varies considerably between species, as does the proportion of the accessory genome designated as strain-specific and variable. As variation is often introduced through horizontal gene transfer, a large proportion of the accessory genome is comprised of mobile genetic elements (MGE) such as conjugative transposons, integrons, and prophages [138,140,374-376]. MGEs are often present as clusters of horizontally acquired genes known as genomic islands which can vary dramatically in size between 10-200 kb [374]. Genes originating from genomic islands often encode for functions involved in mobility, antibiotic resistance, increased survival in novel niches and virulence [377].

Pan-genomes have now been defined for a considerable number of species including *Streptococcus agalactaie, Streptococcus pneumoniae, Escherichia coli, Bacillus cereus, Salmonella Paratyphi A, Haemophilus influenzae and Corynebacterium pseudotuberculosis*, which were all characterised with open pan-genomes, and *Streptococcus pyogenes*, *Ureaplasma urealyticum* and *Bacillus anthracis,* which were classified as closed [139,141,368-370,378,379]. The possession of an open pan-genome indicates that the genetic repertoire of a species can never be fully described, as every new genome will contribute a certain number of novel strain-specific genes, whilst closed a pan-genome denotes a limited degree of genetic variability between strains. Previous pan-genome estimations have also been carried out for both *S. epidermidis* and *S. aureus* using a limited number of strains [361,370]. These studies estimated that *S. aureus* exhibited a closed pan-genome with little intra-species variability whilst the pan-genome of *S. epidermidis* was found to be open.

4.1.4 Horizontal gene transfer

The acquisition and loss of genetic material due to horizontal gene transfer (HGT) plays a significant role in the composition and structure of the accessory genome [380]. The core genome exhibits a relatively low level of genome plasticity and therefore does not experience common or substantial gene rearrangement events, however there is recent evidence that HGT has played a small role in shaping the core gene pool [368].

HGT describes the intra- and inter-cellular exchange of genetic material and is responsible for the prolific spread of antibiotic resistance genes throughout prokaryotic species [381]. In addition to increasing virulence and allowing adaptation to novel environments, HGT can also lead to gene disruptions and deletions that may either increase or decrease fitness. HGT is a result of either conjugation, transduction or transformation of genetic material mediated by a wide variety of mobile genetic elements (MGE), which are primarily responsible for the majority of intraspecies variation[382]. Conjugation requires direct cell to cell contact and involves the transfer of MGEs such as conjugative plasmids and integrated conjugative elements (ICE), which includes conjugative transposons [383]. ICEs are major facilitators of the spread of antibiotic resistance between and within prokaryotic species, existing as clusters of chromosomally located genes which facilitate their own excision and intra- and inter-cellular transfer [383,384]. Transduction is mediated by bacteriophages which act as vectors allowing the integration of phage-transferred genetic material into the host chromosome [385]. In comparison to conjugative and phage-related mechanisms, transformation, which involves the integration of naked DNA from the extracellular environment, usually transfers shorter DNA fragments and requires genetically competent bacteria to act as the donor cell [386].

Genomic islands (GIs) are sporadically distributed clusters of horizontally acquired genes that often account for a large proportion of the observed intra-species variation [387]. They are incorporated into the host chromosome via the transfer of MGEs such as plasmids, conjugative transposons or bacteriophages, and can be substantial in length, often up to 100kb [376]. Pathogenicity islands (PAIs) are a type of GI directly concerned with the pathogenesis of the organism and often encode genes involved with iron uptake systems, toxins, type III secretion systems and adhesins. Other commonly encoded functions within GIs include antibiotic resistance, metal resistance, mobility (integrases and transposases) and secondary metabolites. The phage origin of a large number of GI genes often leads to a large proportion of GI genes classified with unknown function or with no identified homologs[388].

The increase in the size of the pan-genome is a direct result of the incorporation of novel genes through HGT. The massive reservoir of microbial diversity available within different ecosystems has only recently been made apparent with the advent of high-throughput sequencing, and has far surpassed previous estimations made using culture-dependent methods. For example, three outstanding recent studies cataloguing the taxonomic content of the Sargasso Sea, the human gut and soil have estimated the presence of 1,800, 400 and 10^7 microbial species respectively [117,126,389,390]. The extent of novel genes present in the environment was also made apparent by Sargasso Sea study in which 1.2 million previously unknown genes were identified[126].

4.1.5 Methods of *S. epidermidis* strain discrimination

Molecular typing is an essential tool within microbial epidemiology to understand the source and spread of outbreaks and to distinguish clinically important microbial isolates. Molecular typing also has applications in understanding microbial community structure, bacterial etiology and taxonomic identification. Several molecular typing methods have been applied to *S. epidermidis* to understand the distribution and classification of commensal and virulent strains within a clinical setting, and to understand the epidemiological spread of isolates during outbreaks. Some of the more traditional typing methods include pulsed field gel electrophoresis (PFGE), which uses restriction endonucleases to generate strain-specific 'fingerprints', amplified fragment length polymorphisms (AFLP), which follows a similar approach to PFGE but includes amplification of the fragments, PCR ribotyping, which targets intergenic spacer regions to create strain-specific banding patterns, random amplification of polymorphic DNA (RAPD), which uses random 10 bp primers and multi locus sequence typing (MLST), which targets seven essential housekeeping genes [391,394,395,398,401,404].

In recent years, a novel typing technique known as multi-locus variable number tandem repeat analysis (MLVA), has emerged as a increasingly popular technique for epidemiological studies of *S. epidermidis* due to its powerful discriminatory ability, production of inter-laboratory comparable results, high level of reproducibility and low labour and consumable costs [411]. MLVA targets multiple selected loci throughout the genome that are home to tandem repeats flanked by non-repetitive regions using selective primers [408-411]. Visualising the PCR products on an agarose gel allows determination of the band size and subsequent calculation of the number of repeats present at each loci. As MLVA requires variable regions to first be identified using a reference genome and primers designed for the selected loci, a database has been generated which records bacterial tandem repeat patterns (http://minisatellites.u-psud.fr/) as a resource for epidemiological uses [412,413]. This approach has been generated and applied to a

number of species including *Staphylococcus epidermidis*, *Staphylococcus aureus, Escherichia coli* 0157 H7*, Bacillus anthracis, Yersinia pestis, Legionella pneumophila* and *Mycobacterium tuberculosis* [411,414-419].

MLVA has emerged as a increasingly popular technique for epidemiological studies of *S. epidermidis* due to its powerful discriminatory ability, production of intra-laboratory comparable results, high level of reproducibility and low labour and consumable costs [411]. MLVA has been shown to demonstrate a discriminatory power similar to that of PFGE or MLST and has also been shown to be very appropriate for discriminating isolates with high inter-species homogeneity [416,418]. Until recently, PFGE was considered the leading choice for *S. epidermidis* typing due to its high discriminatory ability, however due to the associated high costs and subjective interpretation of banding patterns, MLVA is emerging as a popular alternative [393,420]. Also MLVA has been shown to display a higher degree of discriminatory power than many other typing methods such as PFGE and PCR ribotyping in some species, for example in *S. aureus,* MLVA was able to distinguish isolates classified as identical by PFGE [399,421,422].

4.1.6 Aims of the chapter

S. epidermidis and *S. aureus* represent two of the most important members of the skin microbiome due to their prevalent colonisation of the human body and their potential to cause a myriad of nosocomial-associated infections. Therefore, the main aim of this chapter is to generate a comprehensive description of the intraspecies diversity of *S. epidermidis* and *S. aureus* by generating the pan-genome of both species. To adequately represent the genetic diversity of both species, this study aims to utilise a considerably larger cohort of strains than any previous study by combining the genome sequences of a selection of forearm isolated strains of *S. epidermidis* collected in this study with all publically available *S. epidermidis* genome sequences.

In addition to comprehensively defining the pan-genomes of both *S. epidermidis* and *S. aureus,* this work also aims to generate a detailed comparative analysis between the two species in order to understand the similarities and differences on a species-wide pan-genomic level, rather than the customary single isolate level. Such a detailed species-level comparison will allow the identification of features that define each species, and will allow functional overlaps to be identified. Ultimately, by generating the pan-genomes of both species it will be possible to determine whether the current complement of strains accurately represents the genetic diversity

of the entire species, or whether the sequencing of more strains is required, and will reveal the level of genetic divergence between *S. aureus* and *S. epidermidis*.

4.2 Methods

4.2.1 Isolation of forearm residing staphylococcal strains

To isolate staphylococcal inhabitants of the forearm, nine subjects were sampled using a slightly altered version of the Williamson-Kligman cup-scrub technique [289]. Sampling included placement of a sterile Teflon cup (9.62 cm^2) in the centre of the volar forearm, followed by the application of 1 ml of sampling buffer (0.1 M PBS, pH 8 + 0.1% v/v TWEEN 20) to the area. Using a sterile Teflon stick the skin was gently agitated for 1 min followed by aspiration of the sampling buffer into sterile tubes. To select for *Staphylococcus* spp. only, 0.1 ml of sample was aseptically plated on a staphylococcal specific (SS) media and incubated for 24-48 hrs at 37°C [423] (Appendix A Table 1). Staphylococcal selective components within the SS media include lithium chloride, potassium thiocyanate and sodium azide, whilst the sodium pyruvate acts to counteract certain toxic peroxides which may be produced. The egg yolk component also provides a visual confirmation by causing a zone of clearing to emerge around staphylococcal colonies. Colonies isolated from SS plates were re-plated onto Brain Heart Infusion (BHI) agar to ensure clonal colonies were obtained, and all strains were preserved as duplicate glycerol stocks and stored at -80°C until needed.

4.2.2 DNA extraction

To allow subsequent molecular analysis all suspected staphylococcal strains isolated from the volar forearm were subject to genomic DNA extraction. DNA was extracted using the DNeasy Blood and Tissue kit (QIAGEN), according to manufacturers protocol, with an additional 2 µl of 2.5 ng μl^{-1} lysostaphin added to the initial lysis buffer to ensure complete cell wall lysis [424]. DNA was quantified with the Qubit dsDNA broad range (BR) kit with the Qubit 2.0 spectophotometer (INVITROGEN), and purity was checked with the Nanodrop 2000 (THERMOSCIENTIFIC).

4.2.3 16S rRNA gene amplification and Sanger sequencing

PCR amplification was carried out in a 50 µl reaction mixture containing 25 µl Biomix Red master mix (BIOLINE), 1 µl of forward and reverse primer described in Table 4.1 (10 µM), 2.5 µl of DNA template (10 ng), 19.5 µl ddH20 and 1 µl of high-fidelity enzyme Accuzyme (Table 4.1, BIOLINE). PCR reactions were performed as follows: initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 95°C for 1 min, annealing at 55°C for 1 min and extension at 68°C for 1.5 min. After the last cycle samples were incubated at 68°C for 8 min. The denaturation temperature was calculated according to the manufacturer's instructions,

which took into account the length of the fragment being amplified. PCR amplification products were visualised on a 1.5% agarose gel and quantified using a Qubit dsDNA broad range (BR) kit with the Qubit 2.0 spectophotometer (INVITROGEN). After visualisation PCR products were purified with ExoSAP-IT PCR Clean-up Kit to remove unwanted dNTPs and primers (GE HEALTHCARE).

For each sample a separate 10 μl forward and 10 μl reverse sequencing reaction was prepared by combining 1 μl of ExoSAP cleaned PCR product, 0.25 μl BigDye terminator v3.1 ready reaction mix, 2.25 μl BigDye terminator reaction buffer, 2 μl primer (2 μm) and 4.5 μl molecular grade water, and subjected to a cycle sequencing reaction as follows: 26 cycles of 10 s at 96°C, 5 s at 50°C and 4 min at 60°C, ramping to each temperature at a rate of 1°C per s (APPLIED BIOSCIENCES). Sequencing reactions underwent a clean-up reaction involving the addition of 80 μl precipitation solution comprising: 3 μl of 3 M sodium acetate at pH 4.6, 62.5 μl of non-denatured ethanol and 14.5 μl of deionised water. Samples were precipitated at room temperature for 15 min to remove excess primers. Supernatant was discarded and precipitated samples were washed with 70% ethanol, followed by supernatant removal and re-suspension of sequencing reaction in 14 μl of formamide. Amplified 16S rRNA gene sequences then underwent sequencing on the 3030xl Genetic Analyser (APPLIED BIOSCIENCES).

During the cycling reaction fluorescently labelled dideoxynucleoside triphosphates (ddNTPs) cause termination of the strand extension upon incorporation, due to the absence of a 3-hydroxyl group that prevents the formation of a phosphodiester bond with a subsequent dNTP [425]. This leads to the production of a multitude of different sized fragments for every template, each ending with a fluorescently labelled ddNTP. The ratio of dNTPs to ddNTPs in the BigDye terminator v3.1 ready reaction mix allowed for a termination at least once for every template position, and ensured an even mixture of long and short fragments (APPLIED BIOSCIENCES). Each reaction product was then analysed by capillary electrophoresis using a 3030xl Genetic Analyser, which uses a laser to cause the dyes to fluoresce, each emitting light at a different wavelength (APPLIED BIOSCIENCES). Forward and reverse 16S rRNA gene sequences were quality trimmed and aligned with Codon code aligner (http://www.codoncode.com/aligner/).

Table 4.1. Primers used for 16S rRNA gene sequencing of all suspected staphylococcal strains [243]

Primer Name	Sequence
pF (forward)	AGAGTTTGATCCTGGCTCAG
pR (reverse)	AAGGAGGTGATCCAGCCGCA
pM (middle)	CAGCAGCCGCGGTAATAC

Following the assembly of full-length 16S rRNA gene sequences, taxonomic classifications were generated using the Sequence Match tool of the Ribosomal Database Project (http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp,[426]). This tool identifies the nearest neighbour of each query sequence by calculating matches that contain the highest number of shared 7-mer sequence 'words'. For each pair-wise hit a sequence match score (S_ab score) was calculated by dividing the number of shared 7-mers by the lowest number of total 7-mers, either from the query or RDP sequence. An S_ab score of one signifies a perfect match.

4.2.4 Multi-locus variable-number tandem repeat analysis (MLVA)

MLVA analysis was applied to all identified *S. epidermidis* isolates to allow strain differentiation. As previously described MLVA is a PCR-based typing method which targets five genomic loci home to varying numbers of tandem repeats [411].

For each identified *S. epidermidis* isolate a separate 50 µl PCR reaction was carried out for each of the five selected loci. PCR reaction mixtures comprised 1 µl of each forward and reverse primer (10 µM), 2.5 µl gDNA (10 ng μl^{-1}), 0.5 µl HotStarTaq DNA Polymerase (QIAGEN), 5 µl PCR buffer (QIAGEN), 1 µl dNTP mix (10 mM, LIFE TECHNOLOGIES) and 32 µl molecular grade water. DNA amplification was performed as follows: initial denaturation at 95°C for 5 min, followed by 30 cycles of denaturation at 94°C for 1 min, annealing at the optimum temperature for specific primer pair (Table 4.2) for 1 min and extension at 72°C for 1 min. After the last cycle samples were incubated at 72°C for 10 min. PCR products were visualised on a 1.5% agarose gel and PCR product size was calculated and transformed into repeat copy number for each isolate at each loci. Isolates were classified as different strains if tandem repeat numbers varied at two or more genomic loci.

Table 4.2. Variable-number tandem repeat loci and the flanking primers used to amplify each region [411]. Optimum annealing temperatures were calculated from a varied gradient (data not shown).

Locus	Repeat Motif	Forward Primer	Reverse Primer	Annealing Temperature (°C)
Se1	TCAGACAGCGACTCAGATGGC CCTAGTCCCAACTTGC	GCTGATGGGGAAGAAGTTCA	AACGCTCCTAAACCTGCAAA	55
Se2	TCTGCCTGTTGAATTTCTTTG TGAAATTCTCTTTGTTGG	AGGCCCAAATAAAAAGCAAA	AACTGACGCTCCAGGAGAAG	51
Se3	TCTGAATCACTATCTGAACCC CAACCCCAACTTGCTT	TTTCCGGTATGTGAACCCTTA	TGACACTAGTCGCACAGGAA	55
Se4	TGTCCATGGAATTTCTTCGAA AATTCTCTCTGTTGGGG	TTCATTGTCCCCTGTCTTCT	TCGATCCTGGTAAAGCGATTA	55
Se5	GAATCCGAGTCACTGTCT	GGCCATATAGACCTGGCTTG	AGATGCTGATGGGGAAGATG	57

4.2.5 *S. epidermidis* whole-genome sequencing

Based on the results of the MLVA analysis, five isolates exhibiting different repeat copy-number patterns were selected for whole-genome sequencing. Roche 454 compatible libraries were constructed from 500 ng of purified whole genomic DNA, which was brought to a volume of 100 μl by addition of TE buffer and fragmented by nebulisation for 1 min using 30 psi of nitrogen. Fragmented DNA was purified using the Qiagen MinElute PCR Purification kit (QIAGEN) and eluted into 16 μl TE buffer. Purified samples then underwent end-repair generating blunt-ended DNA fragments followed by an Agencourt AMPure XP (BECKMAN-COULTER) bead purification step to remove excess dNTPs and enzyme. Appropriate MID tags were then ligated to the purified fragments followed by a final Agencourt AMPure XP (BECKMAN-COULTER) bead purification step to remove small fragments. Final libraries were quantified and checked for correct DNA fragmentation by running 1 μl of undiluted sample on a Bioanalyser 2100 using a high sensitivity DNA chip (AGILENT). Additionally 1ul of the final library was quantified using the Qubit dsDNA high sensitivity (HS) kit and measured on the Qubit 2.0 spectophotometer (INVITROGEN). The Centre for Genomic Research (CGR) at the University of Liverpool carried out subsequent emulsion-PCR reactions and 454 Titanium sequencing of all samples.

4.2.6 Read-filtering, assembly and functional annotation

Raw flowgram files were filtered for adaptors using the SFF tools package and fasta and quality files were obtained for each sample (ROCHE). Data was filtered for short and low-quality bases using the NGS QC Toolkit software package using a minimum phred quality score of 20 over at least 70% of the length of the read [427]. Filtered reads were then assembled into contiguous sequences using Roche GD *De Novo* Assembler version 2.6 with the default settings which were as follows: minimum overlap length, 40, minimum overlap identity, 90, alignment identity score, 2 and seed length, 16 (ROCHE).

Assembled contigs were annotated using the freely available online tool IMG [246] .The IMG-ER gene finding protocol begins with a BLAST search against the IMG non-redundant rRNA database to identify rRNA genes, then tRNAScan-SE-1.23, CRISPR recognition tool (CRT) and the Sanger Institute script rfam_scan are used to search for tRNAs, CRISPR sequences and other RNAs respectively [218,244,245]. Protein coding genes are identified using GeneMark [247]. Functional annotation begins with an RPS-BLAST search against the conserved domains

(CDD) and PRIAM databases followed by a filtered hidden markov model search against the Pfam and TIGRfam databases and finally a BLASTp search against the IMG database [248-250,428] .

4.2.7 Pan-genome analysis of *S. epidermidis* and *S. aureus*

4.2.7.1 Genomic data collection

To allow a comprehensive intraspecies analysis of the two important skin-associated species *S. epidermidis* and *S. aureus*, it was necessary to generate the pan-genome from a large collection of strains. Therefore all publically available *S. epidermidis* and *S. aureus* subsp. *aureus* strains with available protein fasta files were downloaded from the National Centre for Biotechnology Information (NCBI) FTP server. Accession IDs and strains IDs are listed in Table 2 and Table 3 of Appendix A.

4.2.7.2 Constructing the pan-genomes

Predicted protein coding genes from all genomes were clustered into orthologous groups using OrthoMCL with the following parameters: e-value cut-off: 1e-5, percentage identity cut-off: 30, percentage match cut-off: 20 [429]. Orthologous gene clusters were defined as core clusters if they contained genes from every genome from that species, and defined as accessory clusters if they contained genes not present in every genome. Singleton genes were defined as those that did not cluster into any orthologous group and were strain-specific.

4.2.7.3 Modelling the core and pan genome

To understand the trend of the pan-genome as additional strains were included, the number of shared genes, total genes and singleton genes were calculated for every number of strains, ie from 2 to 64 strains for *S. epidermidis* and from 2 to 42 strains for *S. aureus*. The order in which strains are added can impact the size of the core and pan genome, however due to the large number of strains used it was not computationally feasible to calculate the size of the core and pan genome for every possible strain combination. Therefore for every number of strains included, the size of the core and pan genome was calculated for 1000 random strain combinations using a bespoke script in the statistical computing language R. To estimate the number of new genes added to the pan-genome upon inclusion of a new strain the number of novel genes was calculated upon addition of every new genome for 1000 different genome

permutations using a similar R script (Appendix B, Table 1. *Pan_core_genome_plot_Rcode, Newly_added_genes_Rcode*).

To calculate the estimated final size of the core genome, an exponential decay model with the equation $n = \kappa * exp(-N/\tau) + tg(\theta)$ was fitted to the core genome plot. *n* represents the number of core genes as a function of the number *N* of genomes, and $tg(\theta)$ represents the predicted size of the core genome. τ , κ and $tg(\theta)$ are free parameters defined to fit the specific curve. This model was also used to estimate the number of novel genes (singletons) that are added to the pan-genome upon the inclusion of additional strains. The state of the pan-genome was predicted using a power law model known as Heaps Law, which can estimate if a pan-genome is open or closed. Heaps Law is defined as follows: $n = \kappa \times N^{\gamma}$, with *n* describing the number of total genes as a function of a number *N of* genomes, and γ and κ describing the slope and intercept of the model respectively. The exponent γ is then used to calculate the following, $\alpha = 1 - \gamma$, which determines whether the pan genome is open ($\alpha < 1$) or closed ($\alpha \geq 1$). When $\alpha \geq 1$ the size of the pan-genome is tending towards a constant value and the addition of novel genomes will not yield novel genes, however when $\alpha < 1$ it indicates the size of the pan-genome will increase as novel genomes are included.

4.2.7.4 Pan-genome annotation

All predicted genes were annotated to a Cluster of Orthologous Groups of proteins (COG) group using WebMGA which utilises RPSBlast to align proteins against the COG database using an e-value cut-off of 0.001 [430]. COG annotations were transferred to the core and accessory clusters and the singletons and the annotation consistency within OrthoMCL clusters was verified using a bespoke perl script (*extract_COGID.pl,* Appendix B, Table 1). COG distributions between *S. aureus* and *S. epidermidis* were compared using a chi-squared test, and enriched COG categories were identified using a Fisher's exact test with false discovery rate correction using MetaStats [431].

4.2.8 Phylogenetic analysis of outlier *S. epidermidis* strain

To investigate the accuracy of the taxonomic classifications of the *S. epidermidis* outlier strain M23864:W1, the 16S rRNA gene sequence was extracted from the annotated genome and aligned against the Ribosomal Database Project (RDP) database using INFERNAL[115,432]. Taxonomy was determined using the RDP classifier with an 80% confidence threshold cut-off and species-level annotation was applied using the RDP SeqMatch tool (previously described) [426]. To investigate and visualise the intra- and inter- species positions of strain M23864:W1 with respect to the *S. epidermidis* strains used in this analysis and a selection of coagulase-negative staphylococci (CoNS), a pan-genome tree was generated for all clustered isolates, and all available CoNS protein sequences. All available draft and complete CoNS genomes were downloaded from the NCBI ftp site and clustered alongside all *S. epidermidis* genomes using OrthoMCL as previously described (Appendix A Table 6). The pan-genome tree was generated by the hierarchical clustering of genomes based on distances calculated from a presence/absence matrix of orthologous gene clusters and strain-specific genes. Each row in the matrix represents a genome whilst each column represents an orthologous gene group or strain-specific gene, with a 1 or 0 representing either the presence or absence of that orthologous group or strain-specific gene in each genome. The distance was calculated by determining the proportion of orthologous groups and strain-specific genes for which the presence/absence status differs, with genomes sharing a larger number of orthologous groups generating a shorter distance than genomes sharing a low number of orthologous groups [122].

4.3 Results and discussion

4.3.1 Isolation of commensal *S. epidermidis* strains

A total of 201 suspected staphylococcal isolates were obtained from the left and right volar forearms of nine volunteers. After 16S rRNA amplification and sequencing, 201 full-length 16S rRNA gene sequences were generated. Taxonomic classification using the RDP database resulted in identification of 57% of sequences to species level and 41% to genus level, with 9% remaining unidentified. Most unclassified isolates generated matches to uncultured bacterium entries.

The staphylococcal species diversity on the volar forearm was found to be quite low with only 1.9 ± 1.3 species identified per sample when isolates classified to the species level were considered. As *S. epidermidis* is considered a ubiquitous species of the skin microbiota, it was no surprise that it was isolated from the majority of samples, and was the predominant staphylococcal species in over 70% of samples (Fig 4.1). Seven coagulase negative staphylococcal (CoNS) species were isolated in total from the forearm along with species from the *Micrococcus, Lysinibacillus* and *Macrococcus* genera (Fig 4.1). Approximately 30% of isolates were either classified to the *Macrococcus* or *Micrococcus* genera. *Macrococcus* is a Gram-positive genus belonging to the family *Staphylococcaceae*, therefore it was unsurprising to find this genus on media selecting for staphylococci. However the presence of a significant number of micrococci was unexpected, as specific components of the media such as sodium azide, potassium thiocyanate, lithium chloride and glycine directly inhibit their growth[423] .

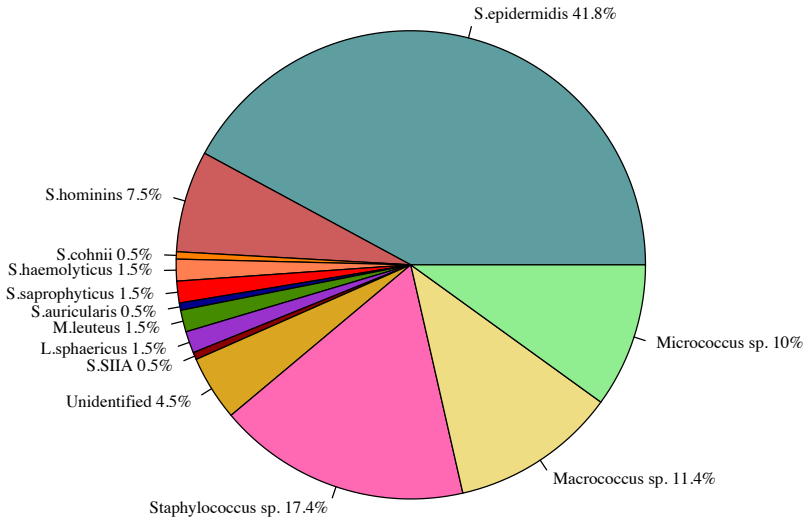


Figure 4.1. Distribution of bacterial species on the volar forearm isolated using staphylococcal specific media and identified using 16S rRNA sequencing. The percentage shown is in relation to the total number of isolates collected.

4.3.1.1 MLVA strain differentiation

A total of 84 isolates were classified as *S. epidermidis,* and were therefore subject to strain differentiation via multi-locus variable-number tandem repeat analysis (MLVA), which targets five loci within the *S. epidermidis* genome that are home to multiple copies of repeats arranged in tandem [411]. No amplification product was generated by any strain for locus site Se5, however the authors who developed the *S. epidermidis* specific protocol also noted that a number of their strains failed to produce amplification product from that specific locus [411]. Three of the five loci, including Se5, were found to be located within open-reading frames responsible for encoding MSCRAMMs belonging to the serine-aspartate (SD) repeat (Sdr) protein family [411]. As well as containing ligand and cell wall binding domains, this family of proteins also contain two repeat regions, the first known as the R region which contains numerous dipeptide SD repeats and the second comprising multiple B-repeats known as the B region [434]. This family of proteins can contain variable numbers of dipeptide SD repeats within the R region without disrupting the open-reading frame. As fibrinogen binding is often associated with pathogenicity via the attachment of microbial cells to extracellular matrix proteins that have coated indwelling

medical devices, it is possible that one or more of members of this protein family are absent in the commensal strains used in this study, explaining the lack of DNA amplification at loci position Se5. The authors who developed this method also noted that the use of four loci only slightly reduced the discriminatory capability in comparison to five, therefore, due to the lack of DNA amplification the use of locus Se5 for further discrimination of these isolates was discarded.

Bands originating from loci 1-4 were successfully amplified in only 42% of isolates, with no amplification from any locus observed for 25% of isolates (Appendix A, Table 5). It was previously demonstrated that DNA is not amplified from these loci in non-*S. epidermidis* strains, suggesting that some of these isolates were incorrectly classified [411]. To ensure distinct isolates were selected for sequencing only isolates from which banding patterns were produced from three or more loci were pursued. This approach led to the removal of 58% of isolates from the collection. As there was no formal threshold for isolate differentiation using MLVA, it was decided for the purposes of this study that any isolate displaying differences in repeat copy number in at least two loci would be classified as a different genotype. In total only five contrasting genotypes were exhibited by isolates (Appendix A, Table 5). This indicates that either MLVA does not exhibit a high enough discriminatory power to differentiate between *S. epidermidis* strains, or demonstrates the clonality of *S. epidermidis* strains isolated from the forearm (Table 4.3).

4.3.1.2 Whole-genome assembly and functional annotation of forearm staphylococci

Following whole genome pyrosequencing of the five selected *S. epidermidis* isolates, high quality reads were assembled into contiguous sequences using the Roche GS *De Novo* Assembler v2.6 (Table 4.4). Excluding sample C20, all assemblies generated 2.4 - 2.5 mbp of contigs, with N50 lengths ranging between 7 - 103 kbp (Table 4.4). Based on an estimated total genome size of 2.5 mb, the assembled contigs of all isolates, except C20, represented at least 96% of the total *S. epidermidis* genome length (Table 4.4). As the assembly of isolate C20 represented a very low proportion of whole genome length, it was excluded from further analysis to avoid biasing of downstream results (Table 4.4). Assembled contigs were submitted to the Integrated Microbial Genomes Expert Review annotation pipeline for structural and functional annotation [246]. Genomes contained an average of 2320 protein coding genes with assigned functional predictions (Table 4.4).

Table 4.3. Copy numbers of tandem repeats at four genomic loci spread throughout the *S. epidermidis* genome. No isolates generated bands from loci Se5 therefore it is not included in this table. Only isolates selected for whole genome sequencing are listed.

	Genomic Locus			
Isolate ID	**Se1**	**Se2**	**Se3**	**Se4**
C20	39	6	28	5
J3	33	9	18	26
M11	41	7	33	7
R1	43	7	30	7
L3	61	7	28	5

Table 4.4. Assembly and annotation statistics of the five *S. epidermidis* forearm isolated strains. *The percentage genome coverage is based on an estimated genome size of 2.5 mb. Contigs were assembled using the Roche GS *De Novo* Assembler v2.6. Functional annotation of contigs was generated using IMG annotation pipeline.

Isolate ID	Total no. Reads	No. Contigs > 500 (bp)	N50 Length (bp)	Percentage Coverage (%)*	Protein Coding Genes	RNA Genes	Protein Coding with Functional Prediction (%)
C20	24,038	987	893	35	n/a	n/a	n/a
J3	140,431	49	82,566	99	2,340	92	83
L3	233,897	70	76,651	97	2,309	92	83
R1	261,375	46	103,907	99	2,351	96	82
M11	297,765	59	81,133	97	2,279	93	83

4.3.2 Pan-genome analysis of *S. epidermidis*

4.3.2.1 Orthologous clustering of all available *S. epidermidis* proteins

To describe the complete pan-genome of *S. epidermidis*, all publically available *S. epidermidis* genome sequences were downloaded from the NCBI FTP server. Addition of the five forearm isolates to this collection generated a dataset of 64 *S. epidermidis* genome sequences representing a diverse mixture of commensal and virulent isolates. All strains exhibited a similar number of predicted protein coding genes with an average of 2,377 per genome. Protein coding genes from all strains in the dataset were then clustered into orthologous groups using OrthoMCL [429]. Clustering a total of 156,877 predicted genes resulted in 4,228 orthologous groups and 6,112 strain-specific or 'singleton' genes. As OrthoMCL identifies recent paralogs and included them in the orthologous groups, 15% of the resulting clusters contained paralogs from one or more of the clustered isolates.

To understand the degree of divergence between each strain the number of singleton genes per strain was calculated. Each clustered genome contained 93 ± 72 strain-specific genes on average, however one strain, *S. epidermidis* M23864:W1 exhibited 360 strain-specific genes, considerably higher than any other *S. epidermidis* isolate (Fig 4.2). For this strain, 14% of the total number of protein coding genes were defined as strain-specific, which was an unusually high proportion compared with the remaining *S. epidermidis* strains. This elevated number of strain-specific genes highlighted the divergence of this strain from the remaining isolates therefore it was removed from the dataset and not included in the clustering analysis. Orthologous clustering of a total of 149,481 predicted genes from the remaining 63 *S. epidermidis* strains resulted in the generation of 4,109 orthologous groups with 10% of clusters containing recent paralogs from one or more of the clustered isolates. The number of singleton genes reduced by 16%, now exhibiting 82 ± 31 strain-specific genes per isolate. In addition, the number of core clusters containing a gene from every strain increased from 764 to 1,377, which is more in agreement with previous estimations that the *S. epidermidis* core genome contains approximately 1,200 genes [361,370]. Due to the considerably larger number of core genes shared by all strains following the exclusion of *S. epidermidis* strain M23864:W1, and the uncharacteristically high number of strain-specific genes, it is possible this strain may have been mis-classified as *S. epidermidis*. Subsequent investigation regarding the initial phylogenetic classifications of the divergent strain was carried out and is discussed in section 4.3.3.

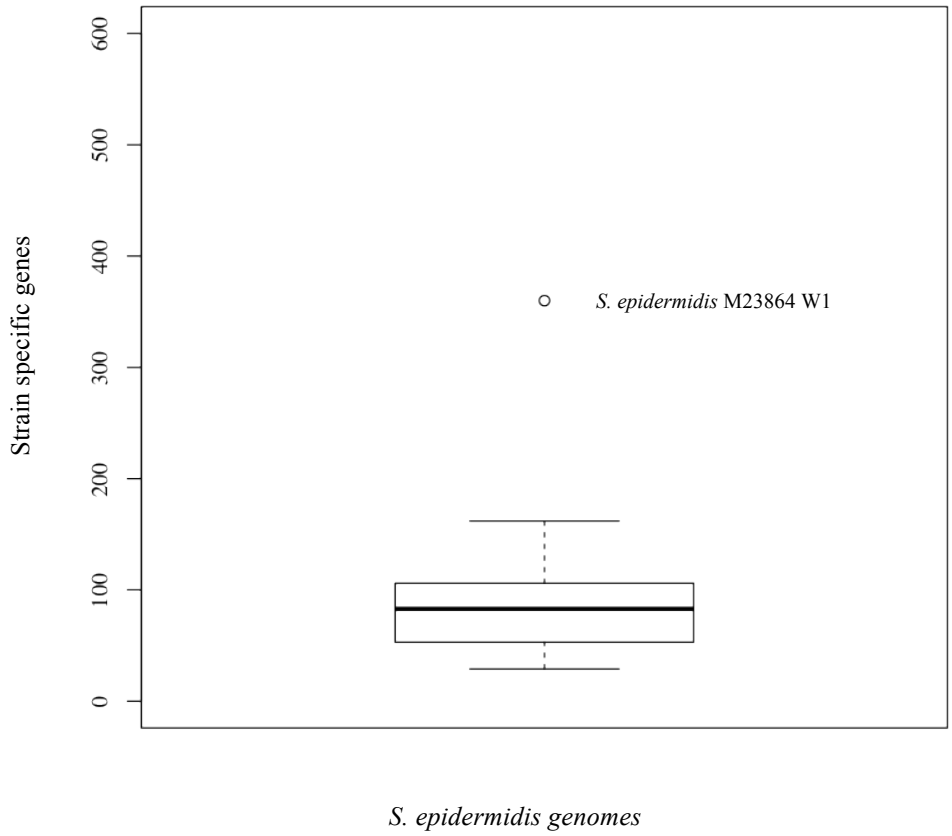


Figure 4.2. The distribution of the number strain-specific genes identified in each strain following orthologous clustering of 64 *S. epidermidis* genomes. The median is shown as a thick black line whilst the perimeters of the box display the 1st and 3rd quantiles of the data. The whiskers extend to the highest and lowest value. The outlier strains: M23864 W1, was excluded from the pan-genome analysis of *S. epidermidis.*

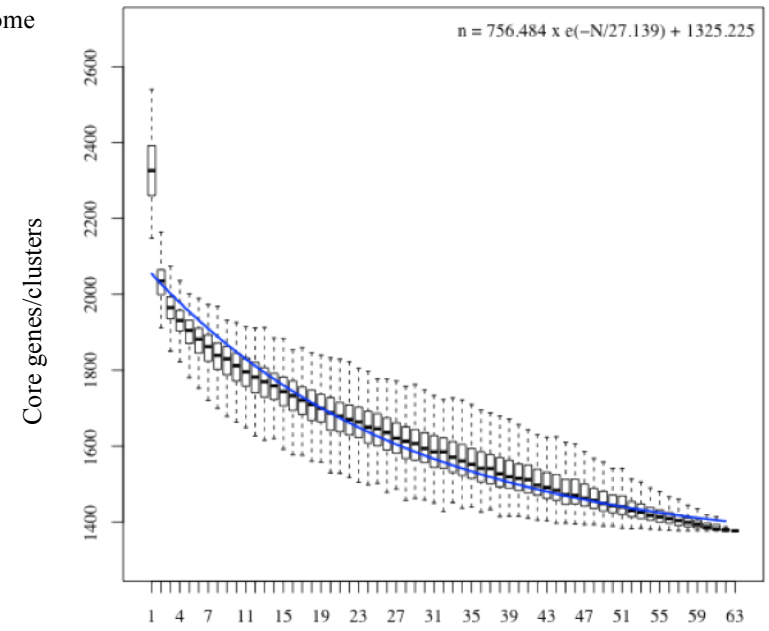
4.3.2.2 Defining the core genome of *S. epidermidis*

To understand the size and structure of the *S. epidermidis* core genome, the number of shared clusters was calculated as a function of the number of genomes. Due to the large number of strains it was computationally unfeasible to calculate the number of shared clusters for all genome combinations, therefore the number of shared clusters was calculated for 1000 random genome combinations for 2-63 genomes (Fig 4.3A). To visualise the trend of the core genome the number of conserved genes was plotted as a function of the number of genomes (Fig 4.3A). The size of the core genome was extrapolated by fitting a non-linear least squares regression to the medians using an exponential decay model which is defined as follows: $n = \kappa * exp(-N/\tau) + tg(\theta)$. n represents the number of core genes as a function of the number N of

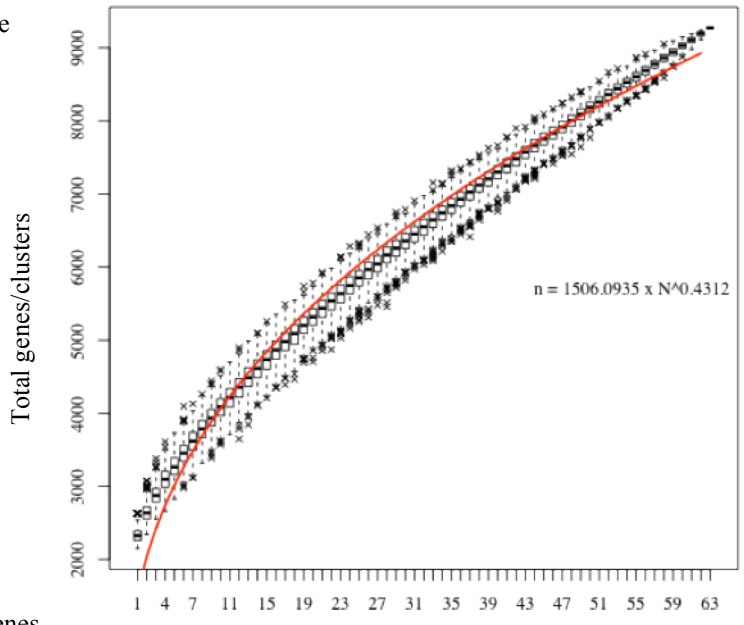
genomes, and *tg(θ)* represents the predicted size of the core genome. τ , κ and tg(θ) are free parameters defined to fit the specific curve. Regression analysis predicted the core genome is approaching an asymptote size of ~1,325 genes (Fig 4.3A). The size of the core genome using 63 *S. epidermidis* strains was found to be ~1,377 genes, indicating the *S. epidermidis* core genome has almost been fully defined by the current genomic data, and sequencing of further strains will not majorly affect the size of the core genome. The majority of the *S. epidermidis* strains used in this analysis were draft genome sequences, so it is possible that the remaining few core genes are present in the un-sequenced regions between the contigs.

The *S. epidermidis* core genome represented a surprisingly small proportion of the entire genetic repertoire of the species, comprising only 15% of the total pan-genome. Within each individual strain an average of 58% of total protein coding genes were classified as core genes, with the remaining 42% of genes drawn from the highly diverse accessory genome. This implies that the majority of the *S. epidermidis* genome is dedicated to core function, and the pan genome is not representative of the gene content within an individual strain. The genome proportion represented by core genes is lower than previously predicted by a recent study, which estimated that 80% of the *S. epidermidis* genome consisted of core genes and only 20% was comprised of accessory genes [361]. As almost three times as many strains were used in this analysis it is not surprising that the size of the core genome has reduced, however it does reiterate the importance of using a large and diverse collection of isolates to define a species pan-genome [361,370]. Although less than previously predicted, core genes still dominate the *S. epidermidis* genome demonstrating that a significant proportion of the coding potential of each strain is dedicated to core functions.

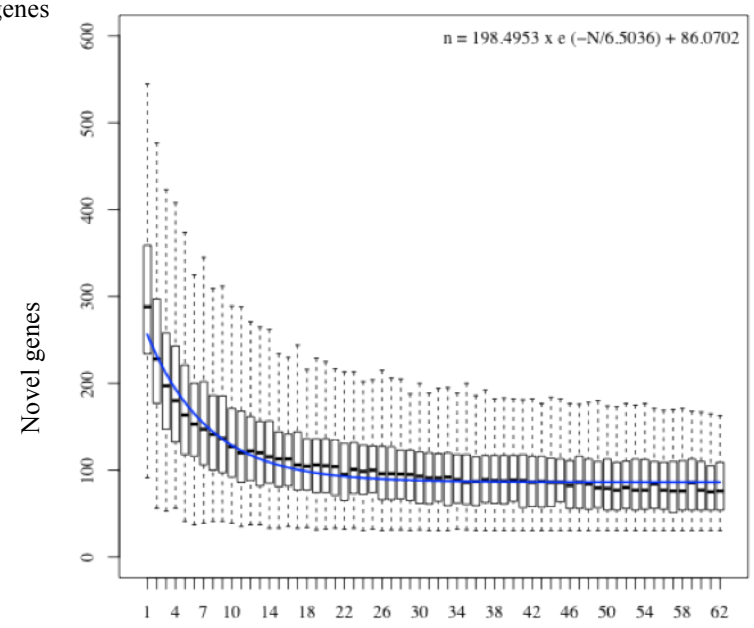
A. Core genome



B. Pan genome



C. Newly added genes



No. of genomes

Figure 4.3. The (A) core genome, (B) pan genome and (C) newly added genes plots for *S. epidermidis*. Boxplots represents the median, 1st and 3rd quartiles, and upper and lower limits of each data point. (A): Each boxplot represents the number *n* of shared genes as a function of the number *N* of genomes. The blue line corresponds to an exponential decay model with the formula $n = \kappa * exp(-N/\tau) + tg(\theta)$ fitted to the median number of shared genes. $tg(\theta)$ predicts the final size of the core-genome. (B): The total number *n* of genes is displayed as a function of the number *N* of genomes as more genomes are sequentially added. The red line represents a non-linear least squares regression fitted with a power law known as Heaps Law, which has the formula n = κ x N^γ. The exponent γ is then used to calculate α that is defined as $1-\gamma$, which determines whether the pan genome is open ($\alpha < 1$) or closed ($\alpha \geq 1$). (C): The number *n* of novel genes added to the pan-genome as more genomes are included. The blue line represents an exponential decay regression model fitted to the median number of new genes, as described for plot A. $tg(\theta)$ represents the number of new genes added to the pan-genome upon the inclusion of a novel genome sequence.

4.3.2.3 The *S. epidermidis* pan genome

The pan genome of a species represents all genes associated with all available strains of that particular species [141]. To understand the total genetic repertoire of *S. epidermidis* the size of the pan-genome was calculated as more genomes were sequentially added. For each genome number *N,* the total number of genes *n* (pan-genome size) was calculated for 1000 random genome combinations to simulate the inclusion of all possible genome combinations.

The size of the *S. epidermidis* pan-genome increased rapidly as more genomes were sequentially included, and when all 63 genomes were considered the final size of the *S. epidermidis* pan-genome concluded at an estimated 9,271 genes, with the exact number of genes varying due to the presence of paralogs (Fig 4.3B). To determine whether the pan-genome was open or closed, a regression analysis was performed using the power law function known as Heaps law. Heaps Law was initially applied to the analysis of languages, and describes the observation that as more text is examined the number of new/unique words found decreases [435]. It was first applied to the analysis of pan-genomes by Tettelin *et al*. in 2008, and has subsequently been used to predict open or closed pan-genomes for a number of organisms [369,370,436-439] Heaps law is defined as follows: n = κ x N^γ, with *n* describing the number of total genes as a function of a number *N of* genomes, and γ and κ describing the slope and intercept of the model respectively. The exponent γ is then used to calculate the following, $\alpha = 1 - \gamma$, which determines whether the pan genome is open ($\alpha < 1$) or closed ($\alpha \geq 1$). When $\alpha \geq 1$ the size of the pan-genome is tending towards a constant value and the addition of novel genomes will not yield novel genes, however when $\alpha < 1$ it indicates the size of the pan-genome will increase as novel genomes are included. Based on the Heaps Law model, regression analysis of the *S. epidermidis* pan-genome resulted in κ = 1506.0935 , γ =

0.4312 and $\alpha = 0.5688$ ($1 - \gamma$), leading to the conclusion that the pan-genome is open and will increase in size upon the addition of novel genomes (Fig 4.3B).

To understand the growth rate of the pan-genome, for each number *N* of genomes the number of *n* new genes added to the pan-genome for 1000 random genome combinations was calculated (Fig 4.3C). The number of novel genes plateaued rapidly as an increasing number of genomes were included, and after the addition of around 11 genomes the number of newly added genes remained quite constant with 96 ± 13 new genes per genome addition, and finalised with an average of 82 ± 31 genes following the addition of the 63rd genome (Fig 4.3C). The number of new genes added per genome was extrapolated by fitting a non-linear least squares regression to the medians using an exponential decay model which is defined as follows: $n = \kappa * exp(-N/\tau) + tg(\theta)$. *n* represents the number of novel genes added as a function of the number *N* of genomes and $tg(\theta)$ represents the rate at which new genes are added to the pan-genome. τ , κ and $tg(\theta)$ are free parameters defined to fit the specific curve. Regression analysis predicted that $tg(\theta) = 86.0702$, estimating the addition of an average of ~86 novel genes to the pan-genome upon the addition of each subsequent genome (Fig 4.3C). This conclusion agrees with the previous estimation of an open pan-genome and demonstrates the large amount of intra-species diversity within *S. epidermidis* [361,370]. The rate of pan-genome expansion of *S. epidermidis* in comparison to other species with open pan-genomes is relatively high, with species such as *S. agalactaie, C. pseudotuberculosis, S. pneumoniae, P. acnes* and *E. coli* exhibiting α values much closer to one and adding a smaller number of average novel genes per new genome [369-371]. Due to the large size of the *S. epidermidis* pan-genome and the small proportion of core genes within the pan-genome, it is likely that further gene additions to the pan-genome will comprise rarer genes that are either strain-specific or present in only a few strains.

4.3.2.4 The *S. epidermidis* accessory genome

As described, core genes comprise only 15% of the *S. epidermidis* pan-genome, with the remaining 85% of genes/clusters defined as dispensable/accessory genes. Considering the large number of strains analysed it was not surprising that singletons dominated a large proportion of the accessory genome, with 5,162 genes classified as strain-specific and the remaining 2,732 genes/clusters shared by at least two but not all strains. A distinct characteristic of *S. epidermidis* is its ubiquitous nature and its ability to colonise a myriad of dry, moist and sebaceous host-associated environments. As this collection of strains represents isolates from multiple environments, it is possible that a considerable number of

the strain-specific genes within the accessory genome are associated with the ability of different strains to colonise and survive within specific niches. These functions would therefore not be conserved throughout the whole species but may appear as core in a sub-selection of strains isolated from the same niche. Therefore although these strain-specific genes are not likely to be involved with functions essential to the basic survival of the species, they may be essential for survival within a specific habitat. However, it is also likely that a large proportion of the accessory genes are not likely to be involved with essential functions and/or represent degenerate functions, and are therefore under a weak selection pressure.

4.3.2.5 Understanding the functional divergence of the *S. epidermidis* core and accessory genome

To understand which functions were enriched within the core and accessory genomes, all clustered proteins were annotated against the COG database. Annotations were subsequently transferred to the corresponding orthologous cluster if all proteins within that cluster generated concurring annotations. Over 50% of accessory clusters and 71% of singleton genes were not assigned a COG annotation, leaving a large proportion of the functional potential of the accessory genome unknown (Fig 4.4). Due to their un-conserved nature, it is likely that a large number of the accessory genes represent novel functions that have not been characterised, and therefore do not contain homologs within the reference databases.

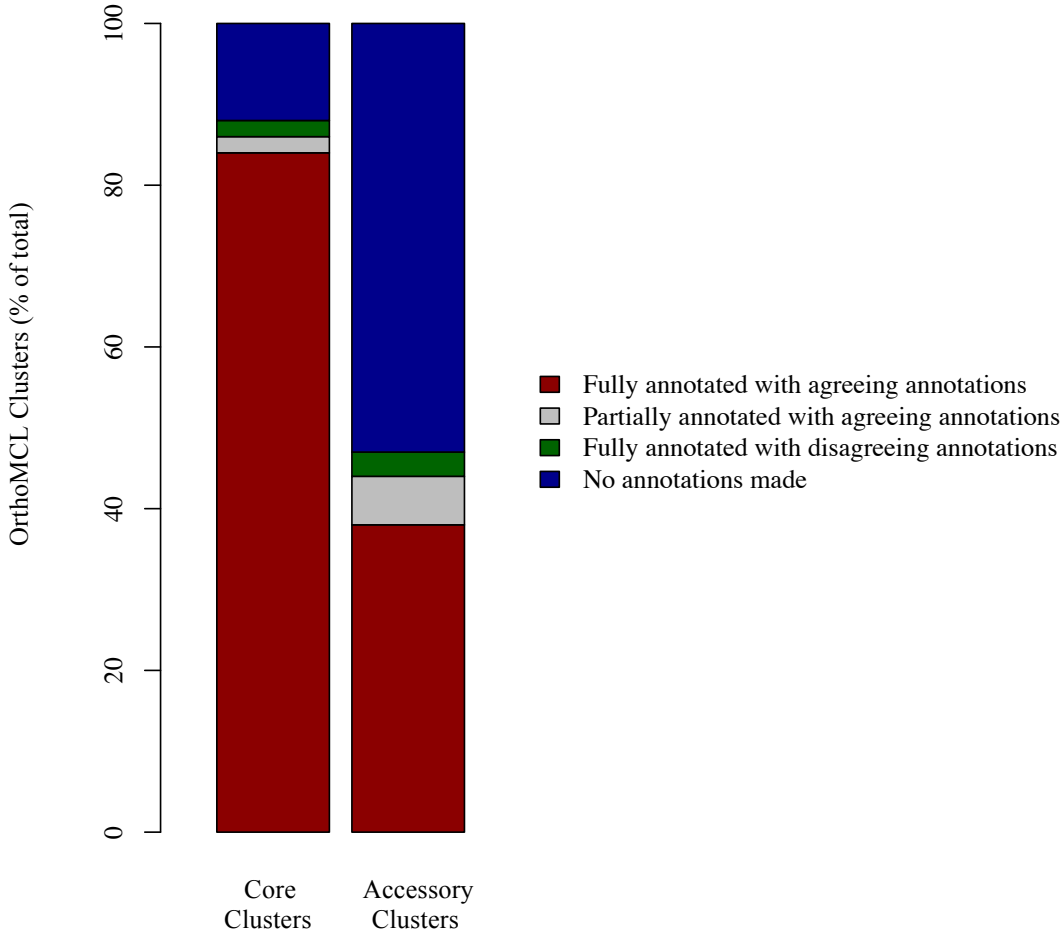


Figure 4.4. Percentage of core and accessory orthologous clusters either A. Fully annotated with agreeing annotations, B. Partially annotated with agreeing annotations, C. Fully annotated with disagreeing annotations or D. No annotations made.

When examining the relative abundance of clusters/genes annotated to each COG category, just over 21% of the core genome, and 8% of the accessory genome, were annotated to poorly characterised COG groups that were either classified with general function prediction only or function unknown categories (Fig 4.5).

When the proportion of genes/clusters annotated to each general COG category were compared between the core and accessory genomes, 17 categories were found to be enriched within the core genome, whilst only two were enriched within the accessory genome (Fig 4.5). Although this indicates the core genome incorporates a greater degree of functional diversity than the accessory genome, it is likely that a considerable proportion of the functional potential of accessory genome is represented by the un-annotated clusters and strain-specific genes, which encompass ~70% of the entire accessory genome.

The two COG categories enriched within the accessory genome were replication, recombination and repair and defence mechanism (Fig 4.5). Over 820 accessory clusters/genes were annotated with functions associated with replication, recombination and repair, however there was relatively little functional diversity within this gene set, with 72% of genes classified as transposase and inactivated derivatives, responsible for the activation and movement of transposable elements. This equates to at least 7% of the entire accessory genome, and 22% of all annotated accessory genes of *S. epidermidis* dedicated to the movement of mobile genetic elements. As transposition is just one method of horizontal gene transfer between strains, this illustrates the extensive arsenal of genes within the *S. epidermidis* accessory genome dedicated to the acquisition of new genetic material, aiding adaptive evolution and acquisition of advantageous genes such as antibiotic resistance. The second enriched COG category within the accessory genome was defence mechanisms (Fig 4.5). The majority of enriched defence genes were associated with resistance to antibiotics and host defences, and included ABC multidrug transporters, antimicrobial peptides (AMP) transporters, beta-lactamase proteins and restriction-modification systems. The enrichment of pathogenicity-associated functions within the accessory genome, and not the core genome, highlights the predominantly commensal lifestyle exhibited by *S. epidermidis*, and demonstrates that the virulent phenotype is limited to subset of strains that have acquired the genes required for infection.

Apart from the poorly characterised COG categories, no other functional category represented more than 2% of the total accessory genes/clusters (Fig 4.5). As the pan-genome of *S. epidermidis* has been characterised as open the genetic diversity of the accessory genome will increase, as subsequent novel strains are included.

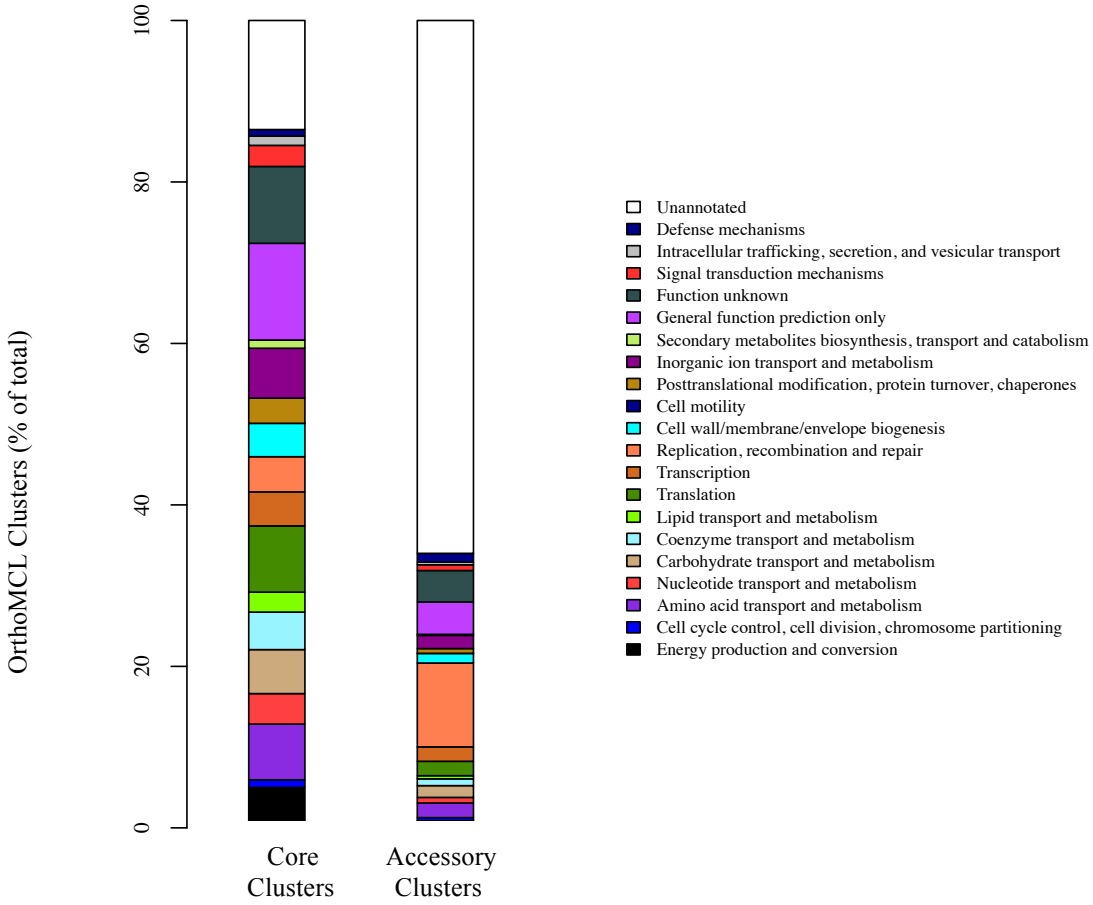


Figure 4.5. The relative abundance of *S. epidermidis* core and accessory genes/clusters assigned to each COG functional category as a function of the total number of clusters.

4.3.3 Phylogenetic analysis of *S. epidermidis* strain M23864:W1

The *S. epidermidis* strain M23864:W1 was not included in the *S. epidermidis* pan-genome analysis due to its divergent gene content and significant impact upon the size of the core genome. To understand if it represents an atypical *S. epidermidis* strain or if it was initially misclassified as *S. epidermidis,* it was subject to an additional phylogenetic classification.

S. epidermidis strain M23864:W1 was isolated from the Department of Paediatric Infectious Diseases in Texas Children's Hospital in 2009, and subsequently underwent whole-genome sequencing (WGS) and annotation to generate a reference genome for the Human Microbiome Project (HMP) (project accession: ACJB00000000 [440]). Phylogenetic

classification was achieved via alignment of the 16S rRNA gene against the SILVA database using ARB [259,441].

To investigate the accuracy of the species-level annotation, the M23864:W1 full length 16S rRNA gene sequence was extracted from the annotated genome and an additional classification was carried out using the SeqMatch tool of RDP. The highest similarity score was generated against an *S. caprae* species, questioning its original taxonomic classification as *S. epidermidis*. Due to recent influx of bacterial reference genomes it is possible that when strain M23864:W1 was originally classified the database did not contain a complete representation of the staphylococcal species, leading to the incorrect classification as *S. epidermidis*. To confirm the phylogenetic classification of strain M23864:W1, the inter-species relationship was investigated by generation of a pan-genome tree (Fig 4.6). This tree was based on hierarchical clustering of the presence/absence profiles of orthologous gene groups within a representative selection of CoNS genomes, as well as all the *S. epidermidis* strains used in this analysis. An increased distance between strains indicates a divergence between the presence/absence profiles. Based on the pan-genome tree the presence/absence profile of *S. epidermidis* strain M23864:W1 was shown to be more similar to profiles of *S. caprae* and *S. capitis* strains than to *S. epidermidis* strains (Fig 4.6). This indicates that the initial phylogenetic classification as *S. epidermidis* was likely to be incorrect, and that this strain should be re-classified.

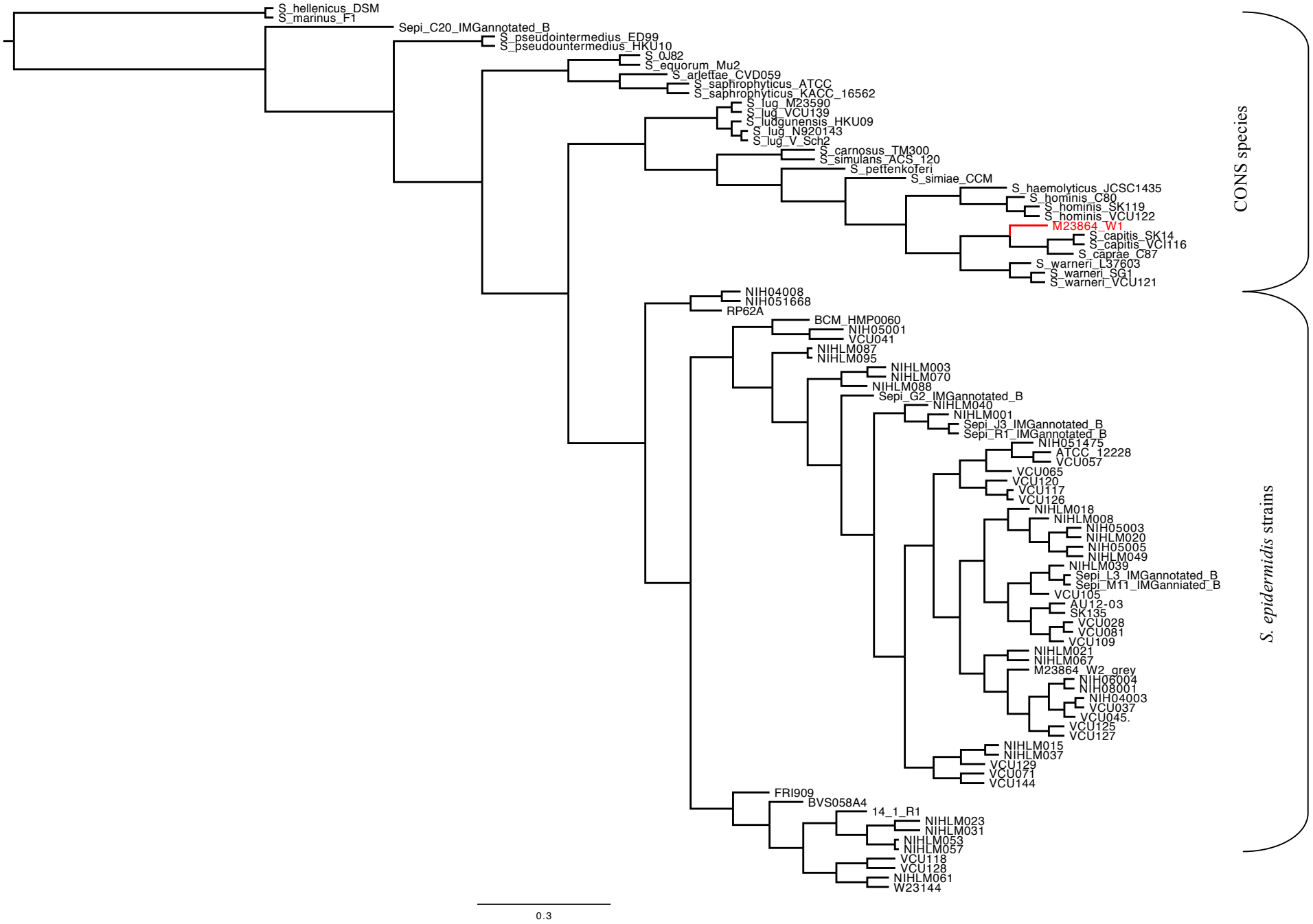

S_hellenicus_DSM
S_marinus_F1
Sepi_C20_IMGannotated_B
S_pseudointermedius_ED99
S_pseudointermedius_HKU10
S_0J82
S_equorum_Mu2
S_arlettae_CVD059
S_saphrophyticus_ATCC
S_saphrophyticus_KACC_16562
S_lug_M23590
S_lug_VCU139
S_ludgunensis_HKU09
S_lug_N920143
S_lug_V_Sch2
S_carnosus_TM300
S_simulans_ACS_120
S_pettenkoferi
S_simiae_CCM
S_haemolyticus_JCSC1435
S_hominis_C80
S_hominis_SK119
S_hominis_VCU122
M23864_W1
S_capitis_SK14
S_capitis_VCI116
S_caprae_C87
S_warneri_L37603
S_warneri_SG1
S_warneri_VCU121
NIH04008
NIH051668
RP62A
BCM_HMP0060
NIH05001
VCU041
NIHLM087
NIHLM095
NIHLM003
NIHLM070
NIHLM088
Sepi_G2_IMGannotated_B
NIHLM040
NIHLM001
Sepi_J3_IMGannotated_B
Sepi_R1_IMGannotated_B
NIH051475
ATCC_12228
VCU057
VCU065
VCU120
VCU117
VCU126
NIHLM018
NIHLM008
NIH05003
NIHLM020
NIH05005
NIHLM049
NIHLM039
Sepi_L3_IMGannotated_B
Sepi_M11_IMGanniated_B
VCU105
AU12-03
SK135
VCU028
VCU081
VCU109
NIHLM021
NIHLM067
M23864_W2_grey
NIH06004
NIH08001
NIH04003
VCU037
VCU045.
VCU125
VCU127
NIHLM015
NIHLM037
VCU129
VCU071
VCU144
FRI909
BVS058A4
14_1_R1
NIHLM023
NIHLM031
NIHLM053
NIHLM057
VCU118
VCU128
NIHLM061
W23144
0.3
CONS species
S. epidermidis strains

Figure 4.6. A pan-genome hierarchical cluster tree generated from a presence/absence matrix of orthologous gene groups within each genome for all *S. epidermidis* strains included in this analysis and a selection of coagulase negative staphylococcal (CONS) strains. The distance between strains is calculated by determining the proportion of orthologous gene groups for which the presence/absence status differs. Strain IDs only are provided for all *S. epidermidis* strains. The strain highlighted in red is the outlier strain M23864:W1. The genus *Staphylococcus* is abbreviated as S.

4.3.4 Comparative analysis of *S. aureus* and *S. epidermidis* pan-genomes

Since *S. aureus* is present in the nasal cavities of ~1/3rd of the population, it is frequently a co-coloniser with the dominant nasal cavity commensal, *S. epidermidis* [59]. Due to the clinical significance of *S. aureus* and the frequent contact and proven interactions with *S. epidermidis*, the pan-genome of *S. aureus* was generated for direct comparison with *S. epidermidis*

4.3.4.1 Generating the *S. aureus* pan genome

All available strains of *S. aureus* subsp. *aureus,* at the time of the analysis, were downloaded from the NCBI ftp server, generating a collection of 42 strains. All protein-coding genes were clustered into orthologous groups using OrthoMCL with parameters as previously described. A total of 112,805 predicted protein-coding genes were clustered into 3,656 orthologous clusters and 2,177 singletons with 7.4% of the clusters containing more than one gene from one or more of the clustered strains.

As seen with *S. epidermidis* the size of the *S. aureus* core genome decreased as more strains were incorporated, however it concluded at a slightly larger size of 1,783 genes/clusters (Fig 4.7A). A lower degree of intra-species variation was exhibited by *S. aureus* in comparison to *S. epidermidis,* with an average of 66% of each individual genome dedicated to core genes, in comparison to the 58% seen in *S. epidermidis* strains. The number of conserved genes *n* was then plotted as a function of the number *N* of genomes by fitting an exponential decay model to the medians as previously described (Fig 4.7A). $tg(\theta)$, which describes the predicted size of the core genome was found to be ~1768 genes (Fig 4.7A). Regression analysis revealed that in agreement with *S. epidermidis*, the *S. aureus* core genome is also nearly completely described and the addition of further strains will not significantly alter its size or content (Fig 4.7A).

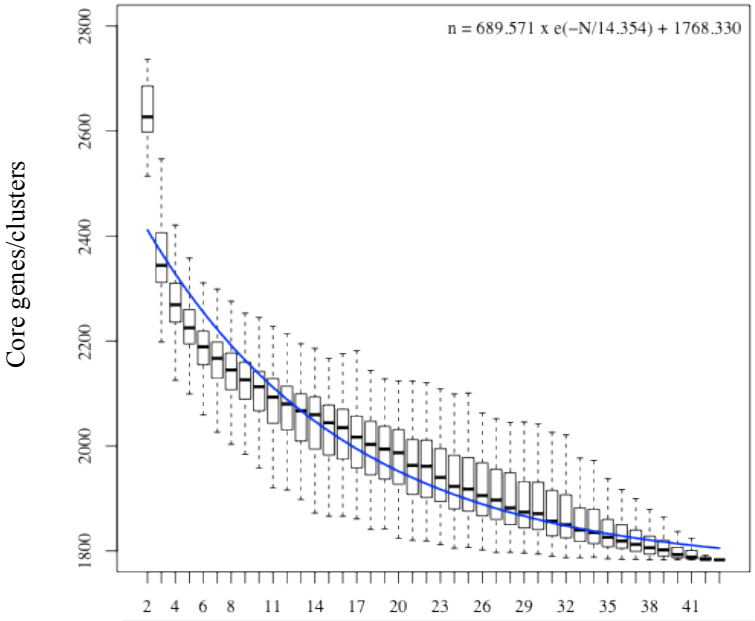
The pan-genome of *S. aureus* increased in size as additional strains were included and concluded at a total size of ~5,833 genes, which included the core genome, genes present in at least two but not all strains and singleton genes (Fig 4.7B). Over 30% of the pan-genome was

occupied by core-genes, double the percentage of the equivalent in the *S. epidermidis* pan genome. The increased proportion of core-genes and correspondingly the reduced proportion of accessory genes demonstrated again the diminished level of intra-species diversity and smaller accessory gene pool in *S. aureus* in comparison to *S. epidermidis*. To understand if the *S. aureus* pan-genome was open or closed a regression analysis was performed on the medians using a power law previously described, known as Heaps Law [435]. This analysis allows calculation of the parameter α, which allowed the pan-genome to then be definitively classified as open ($\alpha < 1$) or closed ($\alpha \geq 1$). Based on the heaps law regression analysis, α was found to be 0.7535 indicating that that *S. aureus* pan-genome is open (Fig 4.7B). This conclusion conflicts with previous estimations that the pan-genome of *S. aureus* is closed, with a predicted α of 1.84, however as approximately 9 strains were used to calculated the pan-genome in that study, it is likely a large proportion of the genetic diversity of the species was missed, resulting in a misleading interpretation of the species pan-genome [370].

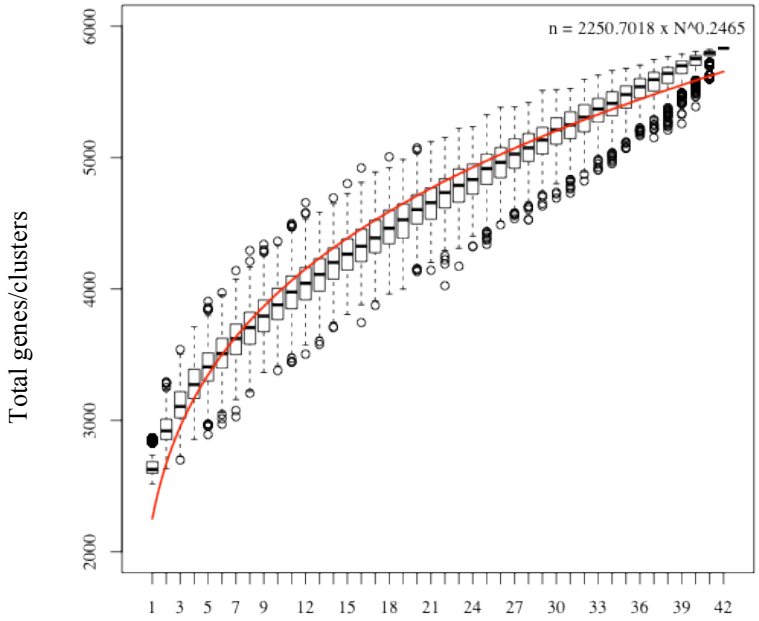
The number *n* of new genes added to the pan-genome for every number *N* of genomes was then calculated as previously described, using 1000 random genome combinations for every *N*. As seen in the corresponding *S. epidermidis* plot, the number of new genes added dropped rapidly as new genomes were included, and plateaued after the addition of ~10 genomes (Fig 4.7C). To calculate the number of new genes added for every subsequent genome included, the data was extrapolated by fitting a non-linear least squares regression to the medians using an exponential decay model as previously described. Regression analysis revealed $tg(\theta)$ to be 41.988, predicting the addition of ~42 genes to the *S. aureus* pan genome for every novel strain (Fig 4.7C).

Although the *S. aureus* pan-genome was classified as open, an α value close to 1 indicates that the pan-genome will experience a reduced level of expansion upon the inclusion of novel genomes in comparison to the *S. epidermidis* pan-genome, which exhibited a smaller α value. The prediction agrees with the considerably smaller total gene repertoire of *S. aureus* in comparison to *S. epidermidis,* which comprised 37% less genes. It could be argued that the reduced total gene repertoire of *S. aureus* is an artefact of the reduced number of strains used to construct the pan-genome in comparison to *S. epidermidis*, however utilising the predicted number of novel genes added per genome to simulate the inclusion of 63 strains, the *S. aureus* pan-genome remaining considerably smaller with ~2,500 fewer genes. It was also determined that the proportion of the *S. aureus* pan-genome dedicated to strain-specific genes was also much smaller with 2,177 (37%) genes/clusters within the pan-genome classified as strain specific, compared to 5,162 (56%) genes/clusters in the *S. epidermidis* pan genome.

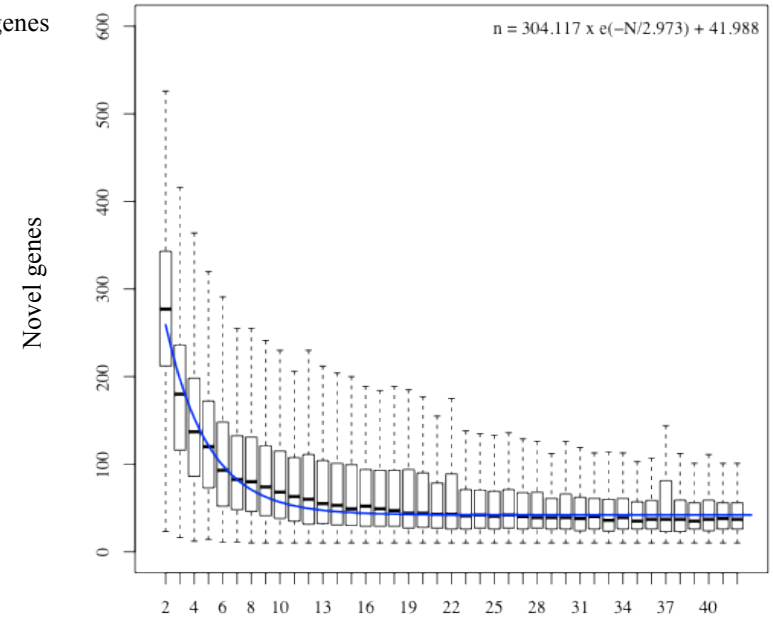
A. Core genome



B. Pan genome



C. Newly added genes



No. of genomes

Figure 4.7. The (A) core genome, (B) pan genome and (C) newly-added genes plots for *S. aureus*. Boxplots represents the median, 1st and 3rd quartiles, and upper and lower limits of each datapoint. (A) Each boxplot represents the number *n* of shared genes as a function of the number *N* of genomes. The blue line corresponds to an exponential decay model with the formula $n = \kappa * exp(-N/\tau) + tg(\theta)$ fitted to the median number of shared genes. $tg(\theta)$ predicts the final size of the pan-genome. (B) The total number *n* of genes is displayed as a function of the number *N* of genomes as more genomes are sequentially added. The red line represents a non-linear least squares regression fitted with a power law known as Heaps Law, which has the formula n = κ x N^γ. The exponent γ is then used to calculate α that is defined as $1-\gamma$, which determines whether the pan genome is open ($\alpha < 1$) or closed ($\alpha \geq 1$). (C) The number *n* of novel genes added to the pan-genome as more genomes are included. The blue line represents an exponential decay regression model fitted to the median number of new genes, as described for plot A. $tg(\theta)$ represents the number of new genes added to the pan-genome upon the inclusion of a novel genome sequence.

4.3.4.2 Comparative functional annotation of the *S. aureus* pan-genome

To understand the functional overlap between the core and accessory genomes of *S. epidermidis* and *S. aureus,* all core clusters, accessory clusters and singleton genes of *S. aureus* were annotated as previously described. The core genomes of *S. epidermidis* and *S. aureus* were shown to be overall functionally quite similar, with no significant difference between the distributions of core gene/clusters over the 25 COG categories ($p > 0.05$, X^2 test). When the abundance of core genes within each individual COG category was compared using Fishers exact test, one category: translation, was found to be significantly enriched within the *S. epidermidis* core genome ($p = < 0.05$), with no significant difference between the abundance of core genes/clusters in all remaining COG categories.

Due to the presence of a large proportion of strain-specific genes within the pan-genomes of both species and the inherent variability of the functions found within the accessory genome it was no surprise that the accessory genomes of *S. epidermidis* and *S. aureus* were found to be much less similar to each other than the core genomes, with significantly different gene/cluster distributions over the 25 functional COG categories ($p < 0.001$, x^2 test). To understand the variation between the functional potential of the *S. aureus* and *S. epidermidis* accessory genomes in more detail Fishers exact test was used to compare the abundance of genes within individual COG categories. Two categories: translation and defence, were found to be significantly enriched within the *S. aureus* accessory genome, whilst only the replication, recombination and repair category was significantly enriched within the *S. epidermidis* accessory genome ($p < 0.001$, $p < 0.05$ and $p < 0.001$ respectively).

4.3.4.3 Identification of a large shared core between *S. epidermidis* and *S. aureus*

To understand the degree of homogeneity between the core genomes of *S. epidermidis* and *S. aureus*, the number of shared core genes and species-specific core genes were determined. A shared core gene was defined as being present in the core genomes of both *S. epidermidis* and *S. aureus*, and a species-specific core gene was defined as being conserved in one species only.

All identified *S. epidermidis* and *S. aureus* core genes were extracted from all strains of both species and clustered into orthologous gene groups using OrthoMCL. Clustering a total of 162,312 core genes resulted in the identification of 1,080 shared core gene clusters present in both the *S. epidermidis* and *S. aureus* core genomes (Fig 4.8). Only 3% of orthologous clusters contained paralogous genes, with one cluster containing a paralog present in all *S.epidermidis* and *S. aureus* species. This shared paralogous cluster was annotated as the membrane associated protein malate-quinone oxoreductase (MQO), which is involved in the tricarboxylic acid (TCA) cycle via the oxidation of malate to oxaloacetate and the subsequent reduction of quinone, which is then utilised by the electron transfer chain (ETC) [442]. The presence of this paralog in both species indicates that the duplication of this gene occurred prior to speciation, and that it has been conserved throughout both individual species, owing to its essential role in aerobic respiration.

Generation of the shared core genome between these two staphylococci demonstrated the high degree of similarity between their reciprocal genetic backbones, and highlighted the large amount of genetic conservation between them (Fig 4.8). There was a larger proportion of shared core genes than species-specific core genes present in the core genomes of both species, as the shared core genome accounted for over 78% and 61% of the individual core genomes of *S. epidermidis* and *S. aureus,* respectively (Fig 4.8). Previous estimations of the size of shared core genome between these two species were based on significantly fewer genomes, and therefore estimated a much larger shared core [359]. Species-specific core orthologous groups were also identified, representing core genes conserved within one species only. *S. aureus* exhibited almost double the number of species-specific core orthologs compared with *S. epidermidis* (Fig 4.8). Due to the reduced intra-species variability within *S. aureus* in relative to *S. epidermidis*, the number of conserved genes between strains is much higher, accounting for the larger core genome.

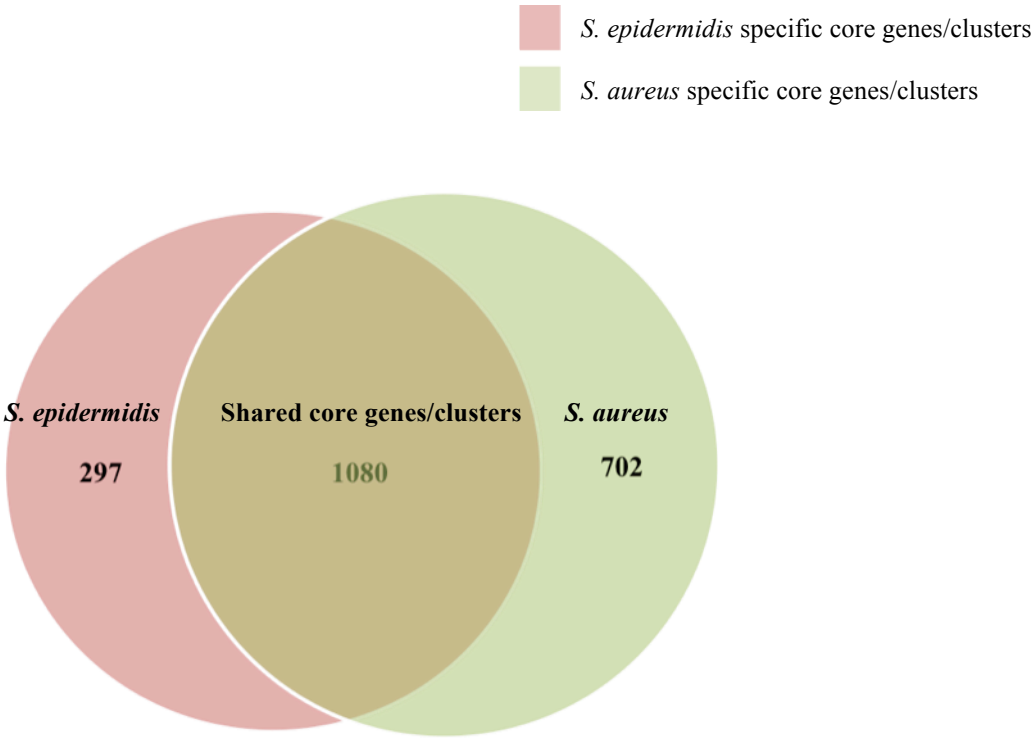


Figure 4.8. Distribution of the number of shared core genes between *S. epidermidis* and *S. aureus* and the number of species-specific core genes.

4.3.4.4 Functional annotation of the species-specific core clusters

To define the functional differences between the species-specific core genomes of *S. aureus* and *S. epidermidis*, all species-specific core orthologous clusters were annotated with a COG group if all genes in that cluster generated matching annotations. Annotations were classified to higher COG categories to identify enriched functions and normalised based on the total number of species-specific clusters. Annotations were directly compared to extract functions unique to the core genome of each species. To ensure the uniquely conserved functions in one species were not present within highly conserved accessory genomes of the alternative species, all species-specific core genes which were present in > 95% of strains of the alternative species were removed from the analysis. At least half of all annotated *S. epidermidis* and *S. aureus* specific core clusters were found to share COG annotations with highly conserved accessory clusters of the opposing species, indicating they represent probable shared functions or genes. Removal of these probable shared core clusters dramatically reduced the size of the *S. aureus* and *S. epidermidis* specific core genomes to approximately 288 and 125 clusters/genes respectively.

A much more diverse set of functions were specific to the *S. aureus* core genome in comparison to the *S. epidermidis* core genome. The majority of *S. aureus* specific core functions were associated with regulation of virulence genes, defence against the host innate immune responses, successful colonisation of the host niche and contribution towards virulence, reflecting the ability of *S. aureus* to cause a wider range of often more serious diseases than *S. epidermidis* which is often more associated with sub acute and chronic infections.

4.3.4.4.1 Regulation of virulence gene expression

Expression of virulence genes in bacteria is a tightly controlled process, mediated by a myriad of transcriptional regulators and two-component regulatory systems (TCSs). TCSs are the main bacterial mechanism by which external environmental signals such as pH level and oxygen concentrations are relayed into the cell leading to expression or repression of specific genes. Due to the larger number of associated virulence genes, it was not surprising that a considerable number of genes classified as specific to the *S. aureus* core genome were involved with the regulation of virulence gene transcription.

The *kdpDE* operon is a TCS, which in *E. coli* is expressed under K^+ limiting conditions and activates the expression of *kdpFABC*, a high affinity kdp-ATPase that acts as an effective K^+ scavenging system [443] . In *S. aureus kdpDE* is not involved in K^+ transportation, as *kdpFABC* is repressed under all K^+ conditions and *kdpDE* is up regulated during biofilm formation in which K^+ is not limited [444]. The role of *kdpDE* in *S. aureus* is thought to be associated with the regulation of virulence factors, as it influences the expression of a variety of genes including the capsular polysaccharide synthetase operon *cap*, the α-toxin gene *hla*, the aureolysin gene *aur,* the lipase gene *geh* and the γ-hemolysin gene *hlgB* [444,445].

The histidine kinase *kdpD* and response regulator *kdpE* were both conserved throughout all *S. aureus* strains and present in the *S. aureus* core genome but not the *S. epidermidis* core genome. The genes *kdpA, kdpB* and *kdpC* within the ATPase encoding operon kdpFABC were also conserved throughout all *S. aureus* strains, however *kdpF* which encodes the small hydrophobic F subunit of the potassium transporter was not conserved between any strains and appeared as un-conserved singleton genes in a small number of strains. In *E. coli kdpF* is not essential for ATPase transporter function *in vivo,* however it acts as a stabiliser of the *kdpABC* complex *in vitro* [446] . Due to the highly conserved nature of the other genes within the kdp operon, it is possible that *kdpF* is not required for ATPase function within *S. aureus*.

The kdp operon has no characterised function in *S. epidermidis*, however orthologs of all kdp genes apart from *kdpF* were identified in 13% of *S. epidermidis* strains.

As maintaining intracellular K^+ levels is essential for cell growth, the possession of an efficient K^+ transporter is an essential function for survival. The *ktr* system has recently been characterised in *S. aureus* as essential for efficient growth in low K^+ environments, and has also been implicated in *S. aureus* pathogenesis [447]. *KtrA, ktrB* and *ktrD* were all found to be highly conserved within *S. aureus* and were therefore classified as core genes, highlighting the importance of this system in K^+ homeostasis. Orthologs of all *ktr* genes were also present in all *S. epidermidis* strains, suggesting the involvement of this system in K^+ uptake in *S. epidermidis* also. As the *kdp* operon is not conserved throughout *S. epidermidis* and an alternative K^+ uptake system has been identified, it is possible that like *S. aureus* the kdp-ATPase does not contribute towards maintaining K^+ homeostasis and the *kdpDE* operon is involved with regulation of gene expression, possibly controlling the transcription of certain virulence genes. It is clear from this analysis that the *kdp* system is a conserved core function in *S. aureus* and an optional accessory function in *S. epidermidis,* possibly contributing towards virulence in clinically isolated strains.

4.3.4.4.2 Iron-acquisition mechanisms

Iron sequestration via high-affinity iron binding proteins such as lactoferrin, transferrin, hemoglobin and ferritin is an effective host defence strategy employed to inhibit the proliferation of invading pathogens [448,449]. As a countermeasure, *S. aureus* has an extensive array of iron-acquisition mechanisms to successfully establish an intra-cellular infection, and responds to iron-limitation by up-regulation of these mechanisms [450,451]. *S. aureus* also uses iron-limitation as a signal of host interaction or disease state to up regulate the transcription of appropriate virulence genes.

The iron regulated surface determinant (Isd) system encompasses nine genes: *IsdA, IsdB, IsdC and IsdH*, which encode proteins covalently bound to the cell wall peptidoglycan, *isdD, isdE* and *isdF,* which comprise an ABC transporter and *isdG* and *isdI, which* encode cytoplasm located enzymes responsible for the degradation of haem [452,453] . Transcription of the *isd* locus has been shown to contribute towards iron acquisition in *S. aureus* via the transport and deconstruction of haem, resulting in the release of free iron within the cell, however it was recently shown that *isdA, isdB* and *isdH* do not effect the ability of *S. aureus* to acquire haem for growth, obscuring the role of these three genes in relation to iron-acquisition [454,455]. All three genes encode pathogenesis associated proteins not related to iron-

acquisition: *isdA* is required for nasal epithelial adhesion and antimicrobial fatty acid resistance, *isdH* is involved in immune evasion and *isdB* encodes a protein with platelet binding function [456-458].

The five i*sd* genes *isdA*, *isdC, isdE, isdF* and *isdG* were present in all *S. aureus* strains and therefore classified as core genes, whilst the remaining four genes, *isdB, isdD, isdH* and *isdI* were conserved within at least 95% of *S. aureus* strains. All *isd* genes apart from *isdG* were unique to the *S. aureus* core genome, and did not appear in the *S. epidermidis* core genome. The presence of the entire *isd* system within core or highly conserved orthologous clusters emphasizes the importance of this system to *S. aureus* proliferation, however whether all *isd* genes are associated with iron-acquisition is unclear [455]. *IsdG* was found to be conserved within *S. epidermidis* as well as *S. aureus,* and encodes a cytoplasmic haem oxygenase responsible for the degradation of haem to biliverdin, CO_2 and free iron in *S. aureus* and *S. lugdunensis* [453] . As all *S. epidermidis* strains contained orthologs of *isdG*, it is possible a similar iron-degradation mechanism is employed by *S. epidermidis.*

A second major mechanism for iron-acquisition in pathogens is the secretion of low molecular weight proteins known as siderophores, which have a high affinity for ferric iron and form iron-siderophore complexes. These are then transported into the cytoplasm by specific transport systems [459]. *S. aureus* is proposed to synthesise four siderophores: staphyloferrin A and B, aurochelin and staphylobactin [460-463].

Production of staphyloferrin B is dependent upon the presence of the sbn nine gene operon, *sbnA-I,* and their inactivation results in decreased growth in iron-limited media [451,464]. The *S. aureus sbn* locus also shares a high sequence similarity to the *sbn* operon of *Ralstonia solanacearum* which has a characterized role in the biosynthesis of staphyloferrin B [465]. In *S. aureus sbnA* and *sbnB* have recently been characterized as L-2,3- diaminopropionic acid (L-Dap) synthetases, which synergistically act to synthesise the L-Dap staphyloferrin B [466]. The sbn operon is important for growth in iron-limited environment and failure to synthesise the siderophore resulted in attenuation of virulence in a murine kidney abscess model [463].

All nine genes within the *sbn* operon are conserved across all *S. aureus* strains indicating the widespread nature and importance of this siderophore biosynthesis mechanism within *S. aureus.* Seven of the nine genes were present in the *S. aureus* specific core genome and the remaining two genes, *sbnA* and *sbnB,* were classified as probable shared core genes due to equivalent annotations applied to two highly conserved *S. epidermidis* accessory genome. Although an identical COG annotation does not imply a corresponding gene, it indicates the

presence of shared protein domains which may indicate a similar biochemical function. Investigation of these *S. epidermidis* conserved genes revealed them to represent an unclassified cysteine synthase and predicted ornithine cyclodeaminase, which have shared homology with the L-Dap synthetases *sbnA* and *sbnB*. The conserved nature of this function within *S. aureus* indicates the essential role of staphyloferrin B during *S. aureus* colonisation and reveals the incorporation of virulence genes into the core genome.

4.3.3.4.3 Capsular polysaccharides

The biosynthesis of extracellular capsular polysaccharides has been extensively characterised in *S. aureus,* and it has been recently shown that *S. epidermidis* can also produce capsular polysaccharide, although the transcription, composition and role of each differs considerably between the two species [467]. The production of extracellular capsular polysaccharides by *S. aureus* aids resistance against host phagocytic killing [468]. Eleven capsular polysaccharide serotypes exist amongst *S. aureus* strains with types 5 and 8 being the most prevalent amongst human-associated isolates [469,470]. The *cap5* and *cap8* operons encode the machinery required for the biosynthesis of capsular polysaccharide type 5 and 8 respectively, and both comprise 16 genes designated *capA-I*. The first seven and last five genes of both loci are extremely similar, and flank type-specific regions comprising *capH, capI, capJ* and *capK* [471]. Accordingly, the capsular polysaccharide types 5 and 8 themselves are biochemically very similar, differing only in the position of the O-acetyl groups and linkages between the amino sugars. Capsule type 5 has the structure →4)-β-D-ManNAcA-(1→4)-α-L-FucNAc(3OAc)-(1→3)-β-D-FucNAc-(1→, whilst type 8 has the structure →3)-β-D-ManNAcA-(4OAc)-(1→3)-α-L-FucNAc-(1→3)-α-D-FucNAc-(1→ [472]. It was recently shown that *S. epidermidis* possesses a *cap* operon capable of encoding the enzymes required to synthesise a poly-γ-DL-glutamic acid (PGA) capsule very similar to the *cap* operon present in *Bacillus* strains [467]. In *S. epidermidis* PGA is a virulence factor involved in survival on the skin and colonisation by conferring resistance to high salt concentrations and innate immune responses such as antimicrobial peptides and neutrophil phagocytosis [467].

Due to the prevalence of capsular polysaccharide types 5 and 8 amongst *S. aureus* strains it was unsurprising to find that all *cap* operon genes apart from the four type-specific genes were highly conserved across all *S. aureus* strains, with orthologs of *capA, capB, capC, capE, capF, capL, capM, capM, capN, capO* and *capP* present in every strain analysed and *capD* and *capG* present in at least 93% of strains. *CapF, capN* and *capP* were the only conserved *cap* genes not unique to the *S. aureus* core genome as they shared annotations with two highly conserved *S. epidermidis* accessory clusters. The *cap5P* gene encodes a UDP-N-

acetylglucosamine (UDP-GlcNAc) 2-epimerase responsible for initiating the biosynthesis of the ManNAcA residue of the capsular polysaccharide type 5 and type 8 by catalyzing the conversion of UDP-GlcNAc to UDP-N-acetyl D-mannosamine (ManNAc) [473] . Recently *cap5P* deficient mutants were shown to still be capable of capsular polysaccharide biosynthesis, despite the vital step catalysed by *cap5P*, likely due to the presence of a UDP-GlcNAc 2-epimerase gene external to the *cap* cluster with functional homology to *cap5P* [473]. Analysis of the highly conserved *S. epidermidis* accessory cluster annotated with the same COGID assigned to the *cap5P* cluster suggested a certain degree of functional similarity, and protein alignment between representative genes from each cluster revealed 77% amino acid similarity. The putative role of UDP-GlcNAc 2-epimerase within *S. epidermidis* is currently unknown. Protein alignments between *capN, capP* and their respective *S. epidermidis* conserved accessory clusters revealed amino acid similarities of 22% and 35% respectively.

The entire *S. epidermidis cap* operon was highly conserved with *capC, capB* and *capD* orthologs present in every strain, and *capA* present in 60 out of 63 strains. The widespread and conserved nature of this operon indicates a strong positive selection within *S. epidermidis* strains.

4.3.3.4.4 Virulence genes

In *S. epidermidis, S. aureus* and other staphylococcal species the *icaADBC* operon encodes enzymes required for the biosynthesis of polysaccharide intercellular (PIA), which mediates cell-cell adhesion and promotes biofilm formation [353,474] . The formation of biofilms following bacterial attachment to implanted medical devices is a major virulence factor for both species and is often associated with increased antibiotic resistance [475,476]Expression of the *ica* operon is subject to environmental regulation and multiple loci/genes have been implicated its transcriptional regulation, including the upstream regulatory gene *icaR,* the stress response factor σ^B , the accessory gene regulator *agr* and the staphylococcal accessory regulator *sarA* [474,477-481].

The entire *ica* operon was found to be highly conserved within *S. aureus* with *icaA, icaB* and *icaR* orthologs present in every strain, and *icaD* and *icaC* orthologs present in over 92% of strains. This degree of conservation classifies the production of PIA as a core function within *S. aureus*, although presence of the operon does not directly correspond to the ability to synthesise PIA. Previous reports have reported the widespread nature of the ica operon within *S. aureus* however there have been some conflicting studies regarding the proportion of strains containing the entire operon [480,482]. The Ica operon was not as highly conserved in *S.*

epidermidis strains in comparison to *S. aureus* with only 30% of strains containing the entire operon and the regulatory gene *icaR*. The teicoplanin-locus regulatory gene *tcaR*, which has also been implicated in the synergistic repression of *icaADBC* expression along with *icaR,* was highly conserved in both *S. aureus* and *S. epidermidis,* present in all *S. aureus* strains and 98% of *S. epidermidis* strains [477,483].

4.4 Conclusion

This research has capitalised upon the recent influx of genomic sequence data to characterise the intraspecies genetic composition of two important nosocomial pathogens, *S. epidermidis* and *S. aureus,* which together are responsible for the majority of hospital-acquired infections in Europe and the USA. By using a larger number of strains than any previous study, it has been possible to more comprehensively describe the pan-genomes of both species than ever before. This work has determined that the core genomes of both *S. epidermidis* and *S. aureus* are almost completely described by the collection of strains used in this analysis, and that sequencing of further strains will not significantly alter the size or gene content of either. The *S. aureus* core genome was found to be larger than the *S. epidermidis* core at ~1,800 genes in comparison to ~1,300 genes. Both pan-genomes were classified as open, however the *S. epidermidis* pan-genome was found to be expanding at a much faster rate with 86 novel genes added to the pan-genome per new strain in comparison to 42 for each new *S. aureus* strain.

This work also revealed that a slightly larger proportion of the *S. aureus* coding potential is dedicated to core function in comparison to *S. epidermidis,* with ~66% of protein coding genes within *S. aureus* genomes classified as core in comparison to ~58% of *S. epidermidis* genes. Using such a large number of *S. epidermidis* genome sequences allowed us to determine that the proportion of the genome dominated by accessory/variable genes is much larger than previously thought, as Conlan *et al*. recently estimated that 20% was dedicated to variable genes, while this research has demonstrated that ~42% of each individual strain comprises variable genes. This indicates a reduced amount of genetic diversity and a higher proportion of conserved function within *S. aureus* strains, which correlates with the increased size of the core genome in comparison to *S. epidermidis*. It also reflects the more specialised lifestyle and niche specificity of *S. aureus,* which frequently colonises just the nasal cavity, in comparison to *S. epidermidis*, which inhabits the majority of host-associated habitats. The diverse lifestyle of *S. epidermidis* suggests it exhibits higher rate of adaptation via novel gene acquisition than *S. aureus*, which is also demonstrated by the considerably larger accessory genome of *S. epidermidis* which concludes at ~7,894 genes in comparison to ~ 4,050 genes.

The functional variation existing between the two species was compared based on the core genome, which is relatively stable for both species. It was found that a considerable degree of the core functions in each species were conserved in both *S. epidermidis* and *S. aureus* strains, despite their divergent lifestyles. Over 60% of the *S. aureus* core genome was also conserved in *S. epidermidis,* and nearly 80% of *S. epidermidis* conserved genes were equivalently classified in *S. aureus*. Previous estimations of the size of shared core genome between these

two species were based on significantly fewer genomes, and therefore estimated a much larger shared core[359]. Functional variation between the core genomes of different species can provide a substantial insight into the basis for species differentiation, therefore core functions specific to each species were extracted and analysed. The functional composition of the specific core-genomes of *S. aureus* and *S. epidermidis* varied quite considerably, with a noted presence of virulence associated genes dominating the *S. aureus* specific core. Functions associated with defence against host responses during an infection such as iron-acquisition mechanisms, protection against innate immune responses including phagocytic killing and regulation of virulence genes were highly conserved throughout the *S. aureus* specific core genome, and absent from the *S. epidermidis* core genome, indicating the high degree of evolutionary conservation pressure placed upon genes allowing a pathogenic lifestyle. This research highlighted the widespread nature of *S. aureus* pathogenicity and classified the presence of a considerable number of virulence-associated genes within the core genome, which has been thought to house primarily housekeeping and essential survival genes. The increased number of virulence factors and genes associated with defence and survival against host innate responses identified in the *S. aureus* specific core genome reflects the ability of *S. aureus* to cause a much wider and more serious range of infections in comparison to *S. epidermidis*.

As more bacterial genome sequences become available it may become more appropriate to re-classify the core genome of a species as comprising the essential core genes, which includes functions essential to survival that are present in all species within a genus or phyla, and the species-specific core, which contains species-specific genes responsible for a diverse range of functions from niche adaptation and survival to virulence. As this study utilised all available strains isolated from a myriad of habitats, it negated the impact of niche-specific adaptations within sub-groups of strains. By generating the pan-genome for groups of strains isolated from the same environment it would be possible to understand the specific functional adaptations required for survival in that niche by identifying sub-group specific conserved genes.

CHAPTER 5

Conclusions and Future Work

This project sought to more comprehensively understand the microbial inhabitants of the human skin surface by fully characterising the axillary microbiota utilising a novel metagenomics approach, and by cataloguing and describing the total intra-species genetic diversity of two important skin-associated species, *S. epidermidis* and *S. aureus*. To ensure the most accurate representation of the microbial community, a number of metagenomic analysis tools were initially subject to an extensive validation process utilising an *in vitro* simulated microbial community.

5.1 Defining the most accurate computation analysis methods for metagenomic data

Generating an accurate description of the taxonomic and functional content of metagenomic datasets is imperative to allow correct predictions to be made regarding the ecological role of the microbial community. The variety of computational tools available for the analysis of metagenomic datasets has led to confusion regarding the most appropriate tool to select in order to precisely reconstruct the microbial community, as different tools often generate contrasting taxonomic and functional profiles. Therefore, to allow the most precise reconstruction of the axillary microbiome in chapter three, the aim of chapter two was to present a comprehensive comparative analysis of popular metagenomic analysis tools, including MetaPhlAn, MG-RAST, IDBA-UD, MetaVelvet and IMG/M, using a simulated *in vitro* microbial community. MetaPhlAn and MG-RAST both considerably overestimated the true species diversity, however by utilising a smaller database of clade-specific marker genes, MetaPhlAn exhibited higher specificity and predicted a more accurate taxonomic profile. Since functional profiling is complicated by the short read length of current sequencing platforms, assembly into contiguous sequences is an important aspect of a metagenomic analysis. Consequently, the accuracy of the assembled contigs can impact the subsequent prediction of community gene content. Utilising IMG/M to annotate MetaVelvet and IDBA-UD assembled contigs revealed that MetaVelvet contigs most comprehensively represented the functional profile of the microbial community, and were more accurate than the corresponding IDBA-UD contigs. By benchmarking numerous tools using an *in vitro* simulated microbial community, it was possible to predict which combination of metagenomic analysis tools would lead to the most precise recreation of the taxonomic and functional content of a microbial community.

5.2 Differential taxonomic profiles generated by whole-genomic data in comparison to 16S rRNA data

This research study presented the first application of whole-genome metagenomic sequencing to a skin-associated microbial community without prior combination of samples or use of a whole-genome amplification technique such as MDA. Utilising a metagenomics approach to characterise the axillary microbiota allowed direct access to the genomic content of the microbial community, and therefore provided an un-biased representation of the taxonomic and functional composition of the axillary microbiome.

Over the last ten years, taxonomic profiling of skin-associated microbial communities has been predominantly achieved using high-throughput 16S rRNA gene sequencing, and although this technique has revolutionised our view of the microbial diversity of the skin, many aspects of the methodology including 16S rRNA copy number variation, PCR-amplification artefacts and primer selection can significantly impact the resulting bacterial diversity estimations and make quantitative approximations and comparative analyses very difficult [302,329,484]. Therefore, it was not surprising to find that the taxonomic profile predicted using the whole-genome shotgun data generated by this study exhibited fundamental differences to the microbiome composition estimated by 16S rRNA gene profiling, namely the absence of *Corynebacterium* as an abundant member of the axillary microbiota [22,52,268,270]. Although all 16S rRNA primers are designed to be universal, preferential amplification or suppression of certain species or taxonomic groups has been reported with the use of certain primer sets, therefore it is possible that *Corynebacterium* is not a dominant member of the axillary microbiome, and its abundance is an artefact of the specific methodology employed [302]. However, as this is the first analysis of the axillary microbiome using whole-genome shotgun sequencing, extensive further work would be required to validate the relative abundance levels of corynebacterial species in the axilla. Also, since few studies have directly compared taxonomic profiles generated using whole-genomic and 16S rRNA data, additional future work is required to comprehensively characterise which method generates the most accurate representation of microbial community composition [328,485].

5.3 The identification of enriched genes within malodorous axillary communities

Microbial inhabitants of the skin have been implicated in a number of host-associated roles ranging from protection against the colonisation of pathogenic bacteria to the etiological factors of a myriad of skin diseases. The axillary microbiota contribute to the generation of malodour via the biotransformation of host-secreted substrates, however the complex

mechanisms by which the majority of malodorous compounds are generated are not fully understood. Extracting the specific biochemical functions or pathways responsible for the generation of axillary malodour, or indeed any skin-associated function, has been limited by the use of 16S rRNA gene profiling-based techniques, which yield taxonomic information only.

Utilising a metagenomic approach to profile high and low axillary malodour microbial communities has allowed the identification of specific microbial genes and functions that may be directly involved in the generation of axillary malodour. Chapter three presented the first use of metagenomics to understand the functional differences between microbial communities associated with high and low axillary malodour. Interestingly, a limited number of functional categories were enriched within microbial communities isolated from high malodour axillae, indicating that a small number of functions are associated with malodour generation, which agrees with the current literature that attributes malodour generation to a distinct group of compounds. None of the over-represented functions within high malodour communities were representative of any genes previously implicated in the generation of malodorous compounds, however as the microbial pathways involved are not completely understood, it is possible that further biochemical characterisation may reveal a potential role of one or more of the enriched functions. This work also revealed a potentially novel mechanism of sulphur-based axillary malodour generation, by the action of two enriched genes, *sufS* and *iscS,* which have been previously characterised as generating malodorous sulphur compounds.

As metagenomic sequencing does not differentiate between live and dead cells, metabolically active and inactive species, and expressed and unexpressed genes, it can only generate hypothetical conclusions regarding the functional activity of the microbial community. To understand the active processes occurring within the axillary microbiome in relation to malodour, additional profiling techniques such as metatranscriptomics, metaproteomics and metabolomics will need to be employed. Metatranscriptomics and metaproteomics allow the characterisation of a microbial community at the transcriptional and translational level respectively, generating profiles of transcriptionally active genes and translated proteins, whilst metabolomics identifies the small metabolite content of a sample. Complementary analyses of microbial communities isolated from high malodour axillae using all three approaches will make it possible to determine the specific cellular processes, metabolic pathways and cellular networks that are involved in axillary malodour generation. Currently, technical challenges such as low sample yield and difficulties associated with protein and metabolite characterisation have prevented the application of these techniques to skin-associated microbial communities, however utilising low-input library preparation kits such

as Nextera is a possible mechanism for metatranscriptomics of low-yield microbial communities in the future.

5.4 An extensive intraspecies diversity is revealed by pan-genome analysis of *S. epidermidis* and *S. aureus*

Although it is important to understand the taxonomic and functional composition of whole microbial communities, it is also relevant to discern the genomic structure of individual species. The influx of microbial genome sequences has revealed the extensive differences in gene content existing between strains of the same species, and therefore it has become apparent that to fully describe a bacterial species it is necessary to understand its pan-genome, which represents the full gene repertoire associated with all characterised strains.

Chapter four presented a complete description and analysis of the pan-genomes of the important skin-associated opportunistic pathogens *S. aureus* and *S. epidermidis*, generating a comprehensive analysis of each species by including more strains than any previous study. The pan-genome analysis of both species revealed the extensive degree in which gene content can vary between strains of the same species. The divergent lifestyles of each species was reflected by the differing pan-genome structures: the more specialised *S. aureus* exhibited a smaller pan-genome with a larger proportion of core functions, whilst the pan-genome of *S. epidermidis* was much larger with an increased proportion of accessory genes, representing its functional ability to colonise a myriad of different niches. A large proportion of the accessory genome of any species is shaped by genes acquired through horizontal gene transfer (HGT) of mobile genetic elements (MGEs) such as transposons or genomic islands, indicating that *S.epidermidis* is subject to the frequent acquisition of novel genetic material.

Although the pan-genomes of *S. aureus* and *S. epidermidis* were distinct in structure, a considerable level of homogeneity was observed between the core genomes, highlighting the conservation of a large number of functions throughout both species. This suggested that although both species exhibit contrasting lifestyles involving different niches and opposing levels of pathogenicity, the core processes which are responsible for their growth and replication characteristics are very similar.

The frequent interactions between *S. aureus* and *S. epidermidis* in host-associated microbial communities generates the opportunity for the frequent exchange of genetic material between the two species, which has already been observed *in vivo* from *S. epidermidis* to *S. aureus*, but

not vice versa [486,487]. The large accessory genome of *S. epidermidis* can therefore be viewed as an evolutionary resource for *S. aureus*, allowing adaptation to new environments or acquisition of novel phylogenetic traits via HGT of MGEs.

In addition to characterising the gene content of an entire species, pan-genome analysis could be applied to sub-sets of strains isolated from distinct niches in order to characterise the genes responsible for adaptation to specific environments. Also, as costs associated with high-throughput sequencing technologies continue to fall, it may become standard place to define a species based on its pan-genome, rather than its single-genome, to reflect the genetic content of the entire species.

5.5 Final conclusion

Work described in this thesis has utilised the significant advances made in DNA sequencing technologies to more comprehensively interrogate the microbial colonisation of the skin. Characterisation of the first skin-associated microbial community has paved the way for the application of metagenomics to a larger range of environments using low-input library preparations kits, and enabling a more complete study of microbial ecology than ever before. To further understand the active processes within microbial communities, the next era of microbial community analysis will comprise techniques such as metatranscriptomics, metaproteomics and metabolomics to ultimately define the ecological roles of specific microbial communities.

APPENDIX A

Appendix A Table 1. Components of the staphylococcal specific (SS) media used to culture forearm isolated strains.

Component	Quantity Added to 1L Deionised Water
Tryptone (g)	10
Lab Lemco Powder (g)	5
Yeast Extract (g)	3
Agar no.1 (g)	13
Sodium Pyruvate (g)	10
Glycine (g)	0.5
Potassium Thiocyanate (g)	22.5
$NaH_2 PO_4.2H_2O$ (g)	1.2
Lithium Chloride (g)	2
Glycerol (ml)	5
Sodium Azide Stock Soln (0.2% w/v) (ml)	10
Sterile Egg Yolk Emulsion (ml)	30

Appendix A Table 2. Strain and accession identifiers of all *S. epidermidis* strains included in the pan-genome analysis.

***S. epidermidis* Strain ID**	**NCBI Accession ID**
Staphylococcus epidermidis ATCC 12228	NC_004461
Staphylococcus epidermidis RP62A	NC_002976
Staphylococcus epidermidis 14.1.R1.SE	NZ_AGUC00000000
Staphylococcus epidermidis AU12-03	NZ_AMCS00000000
Staphylococcus epidermidis BCM-HMP0060	NZ_ACHE00000000
Staphylococcus epidermidis BVS058A4	NZ_AGZV00000000
Staphylococcus epidermidis FRI909	NZ_AENR00000000
Staphylococcus epidermidis M23864:W1	NZ_ACJB00000000
Staphylococcus epidermidis M23864:W2 (grey)	NZ_ADMU00000000
Staphylococcus epidermidis NIH04003	NZ_AKHJ00000000
Staphylococcus epidermidis NIH04008	NZ_AKHF00000000
Staphylococcus epidermidis NIH05001	NZ_AKHE00000000
Staphylococcus epidermidis NIH05003	NZ_AKHI00000000
Staphylococcus epidermidis NIH05005	NZ_AKHD00000000
Staphylococcus epidermidis NIH051475	NZ_AKHL00000000
Staphylococcus epidermidis NIH051668	NZ_AKHK00000000
Staphylococcus epidermidis NIH06004	NZ_AKHH00000000
Staphylococcus epidermidis NIH08001	NZ_AKHG00000000
Staphylococcus epidermidis NIHLM001	NZ_AKHC00000000
Staphylococcus epidermidis NIHLM003	NZ_AKHB00000000
Staphylococcus epidermidis NIHLM008	NZ_AKHA00000000
Staphylococcus epidermidis NIHLM015	NZ_AKGZ00000000
Staphylococcus epidermidis NIHLM018	NZ_AKGY00000000
Staphylococcus epidermidis NIHLM020	NZ_AKGW00000000
Staphylococcus epidermidis NIHLM021	NZ_AKGV00000000
Staphylococcus epidermidis NIHLM023	NZ_AKGU00000000

Appendix A Table 2 cont. Strain and accession identifiers of all *S. epidermidis* strains included in the pan-genome analysis.

***S. epidermidis* Strain ID**	**NCBI Accession ID**
Staphylococcus epidermidis NIHLM031	NZ_AKGX00000000
Staphylococcus epidermidis NIHLM037	NZ_AKGT00000000
Staphylococcus epidermidis NIHLM039	NZ_AKGS00000000
Staphylococcus epidermidis NIHLM040	NZ_AKGR00000000
Staphylococcus epidermidis NIHLM049	NZ_AKGQ00000000
Staphylococcus epidermidis NIHLM053	NZ_AKGP00000000
Staphylococcus epidermidis NIHLM057	NZ_AKGO00000000
Staphylococcus epidermidis NIHLM061	NZ_AKGN00000000
Staphylococcus epidermidis NIHLM067	NZ_AKGM00000000
Staphylococcus epidermidis NIHLM070	NZ_AKGL00000000
Staphylococcus epidermidis NIHLM087	NZ_AKGK00000000
Staphylococcus epidermidis NIHLM088	NZ_AKGJ00000000
Staphylococcus epidermidis NIHLM095	NZ_AKGI00000000
Staphylococcus epidermidis SK135	NZ_ADEY00000000
Staphylococcus epidermidis VCU028	NZ_AFEH00000000
Staphylococcus epidermidis VCU037	NZ_AFTY00000000
Staphylococcus epidermidis VCU041	NZ_AHKX00000000
Staphylococcus epidermidis VCU045	NZ_AFEI00000000
Staphylococcus epidermidis VCU057	NZ_AHKY00000000
Staphylococcus epidermidis VCU065	NZ_AHKZ00000000
Staphylococcus epidermidis VCU071	NZ_AGUB00000000
Staphylococcus epidermidis VCU081	NZ_AHLU00000000
Staphylococcus epidermidis VCU105	NZ_AFTZ00000000
Staphylococcus epidermidis VCU109	NZ_AFUA00000000
Staphylococcus epidermidis VCU117	NZ_AHLA00000000

Appendix A Table 2 cont. Strain and accession identifiers of all *S. epidermidis* strains included in the pan-genome analysis.

***S. epidermidis* Strain ID**	**NCBI Accession ID**
Staphylococcus epidermidis VCU118	NZ_AHLB00000000
Staphylococcus epidermidis VCU120	NZ_AHLC00000000
Staphylococcus epidermidis VCU125	NZ_AHLF00000000
Staphylococcus epidermidis VCU126	NZ_AHLG00000000
Staphylococcus epidermidis VCU127	NZ_AHLH00000000
Staphylococcus epidermidis VCU128	NZ_AHLI00000000
Staphylococcus epidermidis VCU129	NZ_AHLJ00000000
Staphylococcus epidermidis VCU144	NZ_AFED00000000
Staphylococcus epidermidis W23144	NZ_ACJC00000000

Appendix A Table 3. Strain and accession identifiers of all *S. aureus* strains included in the pan-genome analysis.

***Staphylococcus aureus* subsp. *aureus* Strain ID**	**NCBI Accession ID**
Staphylococcus aureus subsp. *aureus* 55:2053	NC_022113
Staphylococcus aureus subsp. *aureus* 58-424	ACUT00000000
Staphylococcus aureus subsp. *aureus* 65-1322	ACJS00000000
Staphylococcus aureus subsp. *aureus* 68.397	ACJT00000000
Staphylococcus aureus subsp. *aureus* A017934:97	ACYP00000000
Staphylococcus aureus subsp. *aureus* Btn1260	ACUU00000000
Staphylococcus aureus subsp. *aureus* C101	ACSP00000000
Staphylococcus aureus subsp. *aureus* C160	ACUV00000000
Staphylococcus aureus subsp. *aureus* CF-Marseille	CABA00000000
Staphylococcus aureus subsp. *aureus* COL	NC_002951
Staphylococcus aureus subsp. *aureus* D139	ACSR00000000
Staphylococcus aureus subsp. *aureus* E1410	ACJU00000000

Appendix A Table 3 cont. Strain and accession identifiers of all *S. aureus* strains included in the pan-genome analysis.

***Staphylococcus aureus* subsp. *aureus* Strain ID**	**NCBI Accession ID**
Staphylococcus aureus subsp. *aureus* ED98	NC_013450.
Staphylococcus aureus subsp. *aureus* EMRSA16	ADAT00000000
Staphylococcus aureus subsp. *aureus* H19	ACSS00000000
Staphylococcus aureus subsp. *aureus* JH1	NC_009632
Staphylococcus aureus subsp. *aureus* JH9	NC_009487
Staphylococcus aureus subsp. *aureus* JKD6008	NC_017341
Staphylococcus aureus subsp. *aureus* JKD6009	ABSA00000000
Staphylococcus aureus subsp. *aureus* M1015	AIYC00000000
Staphylococcus aureus subsp. *aureus* M809	ACUS00000000
Staphylococcus aureus subsp. *aureus* M876	ACJV00000000
Staphylococcus aureus subsp. *aureus* M899	ACSU00000000
Staphylococcus aureus subsp. *aureus* MN8	ACJA00000000
Staphylococcus aureus subsp. *aureus* MRSA252	NC_002952
Staphylococcus aureus subsp. *aureus* EMRSA16	NC_002953
Staphylococcus aureus subsp. *aureus* MSSA476	NC_003923
Staphylococcus aureus subsp. *aureus* MW2	NC_009782
Staphylococcus aureus subsp. *aureus* Mu3	BABM00000000
Staphylococcus aureus subsp. *aureus* Mu50-omega	C_002758
Staphylococcus aureus subsp. *aureus* Mu50	NC_002745
Staphylococcus aureus subsp. *aureus* N315	NC_007795
Staphylococcus aureus subsp. *aureus* NCTC8325	ACHD00000000
Staphylococcus aureus subsp. *aureus* TCH130	NC_017342
Staphylococcus aureus subsp. *aureus* TCH60	ACHH00000000
Staphylococcus aureus subsp. *aureus* USA300-FPR3757	NC_007793
Staphylococcus aureus subsp. *aureus* USA300-TCH959	NC_010079
Staphylococcus aureus subsp. *aureus* USA300_TCH1516	NC_010079
Staphylococcus aureus subsp. *aureus* WBG10049	ACSV00000000

Appendix A Table 3 cont. Strain and accession identifiers of all *S. aureus* strains included in the pan-genome analysis.

***Staphylococcus aureus* subsp. *aureus* Strain ID**	**NCBI Accession ID**
Staphylococcus aureus subsp. *aureus* WW2703:97	ACSW00000000
Staphylococcus aureus subsp. *aureus* c427	ACSQ00000000
Staphylococcus aureus subsp. *aureus* newman	NC_009641

Appendix A Table 4. Strain and accession identifiers for all CONS staphylococcal strains used to generated the staphylococcal pan-genome tree to investigate the phylogenetic position of *S. epidermidis* outlier strains

Species ID	**NCBI Accession ID**
Staphylococcus carnosus TM300	NC_012121
Staphylococcus haemolyticus JCSC1435	NC_007168
Staphylococcus lugdunensis HKU09 01	NC_013893
Staphylococcus lugdunensis N920143	NC_017353
Staphylococcus saprophyticus ATCC 15305	NC_007350
Staphylococcus warneri SG1	NC_020164
Staphylococcus equorum Mu2	NZ_CAJL00000000
Staphylococcus hominis C80	NZ_ACRM00000000
Staphylococcus hominis SK119	NZ_ACLP00000000
Staphylococcus hominis VCU122	NZ_AHLD00000000
Staphylococcus lugdunensis ACS 027 V Sch2	NZ_AGZW00000000
Staphylococcus lugdunensis M23590	NZ_AEQA00000000
Staphylococcus lugdunensis VCU139	NZ_AHLK00000000
Staphylococcus OJ82	NZ_ALPU00000000
Staphylococcus pettenkoferi VCU012	NZ_AGUA00000000
Staphylococcus saprophyticus KACC 16562	NZ_AHKB00000000
Staphylococcus simiae CCM 7213	NZ_AEUN00000000
Staphylococcus simulans ACS 120 V Sch1	NZ_AGZX00000000
Staphylococcus warneri L37603	NZ_ACPZ00000000
Staphylococcus warneri VCU121	NZ_AFEC00000000

Appendix A Table 5. Copy numbers of tandem repeats at four genomic loci spread throughout the *S.epidermidis* genome. No isolates generated bands from loci Se5 therefore it is not included in this table. Only isolates generating bands from 3 or more loci are included. The isolates selected for whole genome sequencing appear in bold.

	Genomic loci			
Isolate ID	**Se1**	**Se2**	**Se3**	**Se4**
A4	17	8	25	4
A5	17	8	25	4
A19	41	8	10	5
B9	22	8	32	4
B17	39	8	29	5
B19	19	8	10	5
C13	39	6	33	5
C15	28	8	28	5
C20	39	6	28	5
E1	23	no band	23	4
E2	23	no band	23	7
E4	28	no band	28	4
E6	23	no band	23	4
F1	23	no band	23	4
G2	21	no band	no band	band
H1	35	no band	26	7
H3	28	no band	23	5
I8	19	8	18	6
I17	53	7	18	6
I22	41	10	10	5
J3	33	9	18	26
J19	38	8	10	5
J29	44	7	13	6
J33	38	6	10	5
J36	31	no band	27	18

Appendix A Table 5 cont. Copy numbers of tandem repeats at four genomic loci spread throughout the *S.epidermidis* genome. No isolates generated bands from loci Se5 therefore it is not included in this table. Only isolates generating bands from 3 or more loci are included. The isolates selected for whole genome sequencing appear in bold.

	Genomic loci			
Isolate ID	**Se1**	**Se2**	**Se3**	**Se4**
J40	50	8	18	4
L3	61	7	28	5
L9	39	8	28	5
M2	20	7	23	7
M11	41	7	33	7
R1	43	7	30	7
R2	55	6	28	5
R3	43	7	30	7

Appendix A Table 6. Strain and accession identifiers for all CONS staphylococcal strains used to generated the staphylococcal pan-genome tree to investigate the phylogenetic position of *S. epidermidis* outlier strains

Species ID	NCBI Accession ID
Staphylococcus carnosus TM300	NC_012121
Staphylococcus haemolyticus JCSC1435	NC_007168
Staphylococcus lugdunensis HKU09 01	NC_013893
Staphylococcus lugdunensis N920143	NC_017353
Staphylococcus saprophyticus ATCC 15305	NC_007350
Staphylococcus warneri SG1	NC_020164
Staphylococcus equorum Mu2	NZ_CAJL00000000
Staphylococcus hominis C80	NZ_ACRM00000000
Staphylococcus hominis SK119	NZ_ACLP00000000
Staphylococcus hominis VCU122	NZ_AHLD00000000
Staphylococcus lugdunensis ACS 027 V Sch2	NZ_AGZW00000000
Staphylococcus lugdunensis M23590	NZ_AEQA00000000
Staphylococcus lugdunensis VCU139	NZ_AHLK00000000
Staphylococcus OJ82	NZ_ALPU00000000
Staphylococcus pettenkoferi VCU012	NZ_AGUA00000000
Staphylococcus saprophyticus KACC 16562	NZ_AHKB00000000
Staphylococcus simiae CCM 7213	NZ_AEUN00000000
Staphylococcus simulans ACS 120 V Sch1	NZ_AGZX00000000
Staphylococcus warneri L37603	NZ_ACPZ00000000
Staphylococcus warneri VCU121	NZ_AFEC00000000

APPENDIX B

Appendix B Table 1. Description and usage instructions for all bespoke programs/scripts used in this thesis. All scripts are freely available and provided in the attached DVD.

Referred page in thesis	Name of program/script	Author	Description	Usage
36	blasttrim	Richard Gregory	Removes 19 bp transposon sequence from 454 and Ion Torrent reads (fasta format only)	*blasttrim -i reads -a transposon.fasta -o trimmed_reads* **Input files:** ➢ *reads* : prefix of fasta and optional qual file ➢ *transposon.fasta* : fasta file of transposon sequence **Output files:** ➢ *trimmed_reads* : fasta file of trimmed reads
37	mergeRead_fastq.pl	Jennifer Kelly	Merges paired end reads for subsequent artificial duplicate removal using PrinSeq (fastq format only).	*mergeRead_fastq.pl forward.fastq reverse.fastq > mergedreads.fastq* **Input files:** ➢ *forward.fastq* : fastq file of forward (R1) reads ➢ *reverse.fastq* : fastq file of reverse (R2) reads **Output files:** ➢ *mergedreads.fastq* : fastq file containing appended R1 and R2 sequence and quality strings

37	splitRead_fastq.pl	Jennifer Kelly	Splits merged de-replicated reads back into forward and reverse read files following de-replication by PrinSeq.	*splitRead_fastq.pl dereplicated_reads.fastq* **Output files:** ➢ *read1.fastq* : forward reads ➢ *read2.fastq* : reverse reads
37	extract_full_partial.pl	Jennifer Kelly	Extracts the number of fully and partially mapped reads to each reference genome. Only applicable for Newbler mappings.	*extract_full_partial.pl –i 454ReadStatus.txt > mapping_stats* **Input file:** ➢ *454ReadStatus.txt* : standard file generated by Newbler following mapping **Output file**: ➢ *mapping_stats* : tab-delimited file containing number of partially and fully mapped reads per reference
37	extract_coverage.pl	Jennifer Kelly	Calculates total contig length and average coverage from mapping. Only applicable for Newbler mappings.	*extract_coverage.pl 454AllContigs.fna 454AlignmentInfo.tsv > coverage_stats* **Input files:** ➢ *454AllContigs.fna* : standard file generated by Newbler following mapping ➢ *454AlignmentInfo.tsv* : standard file generated by Newbler following mapping **Output file:** ➢ *coverage_stats* : tab-delimited file containing total contig length and average coverage for each reference
37	coverageStatsSplitByChr_v2.pl	Kevin Ashelford	Calculates coverage statistics for mapped reads.	*coverageStatsSplitByChr_v2.pl –i mappedfile.bam > coverage_stats*

			Only applicable to BAM formatted files.	**Input file:** ➢ *mappedfile.bam* : BAM file of mapped reads generated by BWA **Output file**: ➢ *mapping_stats* : tab-delimited file containing the following 13 columns for each reference: 1: Reference ID 2: Length of reference (bp) 3: Length of mapped reference (bp) 4: % of reference mapped to 5: The mean coverage of the mapping 6: The median coverage of the mapping 7: The standard deviation of the coverage 8 : Quartile 1 coverage 9: Quartile 3 coverage 10: 2.5% coverage 11: 97.5% coverage 12: min coverage 13: max coverage
38	chimericity_of_assembly.sh	Jennifer Kelly	Bash script containing the pipeline used to identify chimeric and non-chimeric contigs, and the degree of chimericity of each chimeric contigs.	* Refer to bash script for specific usage and input file details
38	fixed_chimericity_of_assembly_simple.pl	Jennifer Kelly	Perl script utilised in chimericity_of_assembly.sh pipeline.	*fixed_chimericity_of_assembly_simple.pl -R refmapping.out –C contigmapping.out* **Input files:** ➢ *refmapping.out*: tab-delimited file containing the reference ID of each

				read ➢ *contigmapping.out*: tab-delimited file containing the contig ID of each read **Output files:** ➢ *degreeChimericity.out* : list of chimeric contigs & the degree of chimericity of each contig. ➢ *chimeric_contigs.out* : the reference ID of each read within each chimeric contig ➢ *chimeric_nonchimeric_contigs.out* : the chimeric status of each contig (chimeric/non-chimeric/non-chimeric-unID).
125	Pan_core_genome_plot_Rcode.rdata	Ben Wareham	R code used to calculate core and pan genome sizes for 1000 strain combinations for increasing numbers of strains. Code is loaded into R, and then the function 'Wrapper' is used to calculate core/pan genome sizes.	*Wrapper(data,1000) -> pancore.csv* **Input file:** ➢ *data*: tab-delimited file presence/absence (1/0) of each orthologous cluster within each strain **Output file:** ➢ *pancore.csv*: comma-separated file containing three columns for each N strains: col1 = IDs of selected strains, col2 = size of core genome, col3 = size of pan genome.
125	Newly_added_genes_Rcode.rdata	Ben Wareham	R code used to calculate number of newly added genes for 1000 strain combinations for increasing	*Wrapper(data,1000) -> newgenes.csv* **Input file:** ➢ *data*: tab-delimited file

			numbers of strains.	presence/absence (1/0) of each orthologous cluster within each strain **Output file:** ➢ *pancore.csv*: comma-separated file containing two columns for each N strains: col1 = IDs of selected strains, col2 = number of newly added genes
125	*extract_COGID.pl*	Jennifer Kelly	Adds COG annotations to each orthologous cluster if all annotations within cluster concur.	*extact_COGID.pl all_orthomcl.out COGIDlist.txt > annotated_clusters.txt* **Input files:** ➢ *all_orthomcl.out*: output file from orthologous clustering using OrthoMCL. ➢ *COGIDlist*.txt : tab-delimited file of all clustered genes and the corresponding COGID (obtained from http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/cog/) **Output file:** ➢ *pancore.csv*: comma-separated file containing three columns for each N strains: col1 = IDs of selected strains, col2 = size of core genome, col3 = size of pan genome.

REFERENCES

1. Nicoll, P. & Cortese, T. THE PHYSIOLOGY OF SKIN. *Annual Review of Physiology* 177–203 (1972).
2. Allen, T. D. & Potten, C. S. Fine-structural identification and organization of the epidermal proliferative unit. *Journal of cell science* **15,** 291–319 (1974).
3. Iizuka, H. Epidermal turnover time. *Journal of Dermatological Science* **8,** 215–217 (1994).
4. Sato, K., Leidal, R. & Sato, F. Morphology and development of an apoeccrine sweat gland in human axillae. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **252,** R166–R180 (1987).
5. Sato, K. & Sato, F. Sweat secretion by human axillary apoeccrine sweat gland in vitro. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **252,** R181–R187 (1987).
6. Andrews, G. C., Domonkos, A. N., Arnold, H. L. & Odom, R. B. Andrews' Diseases of the skin: Clinical dermatology. *Journal of the Royal Society of Medicine* **76** (1983).
7. Zouboulis, C. C. Acne and sebaceous gland function. *Clinics in Dermatology* **22,** 360–366 (2004).
8. Nikkari, T. Comparative chemistry of sebum. *Journal of Investigative Dermatology* **62,** 257–267 (1974).
9. Roth, R. R. & James, W. D. Microbial ecology of the skin. *Annual Reviews in Microbiology* **42,** 441–464 (1988).
10. Dikstein, S. & Zlotogorski, A. Measurement of skin pH. *Acta Derm Venereol Suppl (Stockh)* **185,** 18–20 (1994).
11. Ehlers, C., Ivens, U. I., Møller, M. L., Senderovitz, T. & Serup, J. Comparison of two pH meters used for skin surface pH measurement: the pH meter 'pH900'from Courage & Khazaka versus the pH meter "1140"from Mettler Toledo. *Skin research and Technology* **7,** 84–89 (2001).
12. Schmid-Wendtner, M. H. & Korting, H. C. The pH of the Skin Surface and Its Impact on the Barrier Function. *Skin Pharmacol Physiol* **19,** 296–302 (2006).
13. Rieg, S., Garbe, C., Sauer, B., Kalbacher, H. & Schittek, B. Dermcidin is constitutively produced by eccrine sweat glands and is not induced in epidermal cells under inflammatory skin conditions. *British Journal of Dermatology* **151,** 534–539 (2004).
14. Schittek, B. *et al.* Dermicidin: a novel human antimicrobial peptide secreted by sweat glands. *Nature Immunology* **2,** 1133–1137 (2001).
15. Lai, Y.-P. *et al.* Functional and structural characterisation of recombinant dermcidin-1L, a human antimicrobial peptide. *Biochemical and Biophysical Research Communications* **328,** 243–250 (2005).
16. White, S. H., Wimley, W. C. & Selsted, M. E. Structure, function, and membrane integration of defensins. *Current opinion in structural biology* **5,** 521–527 (1995).
17. Harder, J., Meyer-Hoffert, U., Wehkamp, K., Schwichtenberg, L. & Schröder, J.-M. Differential gene induction of human β-defensins (hBD-1,-2,-3, and-4) in keratinocytes is inhibited by retinoic acid. *Journal of Investigative Dermatology* **123,** 522–529 (2004).
18. Fredricks, D. N. Microbial ecology of human skin in health and disease. *Journal of Investigative Dermatology Symposium Proceedings* **6,** 167–169 (2002).
20. Wilson, M. *Microbial Inhabitants of Humans*. (Cambridge University Press, 2005).
21. Grice, E. A. *et al.* A diversity profile of the human skin microbiota. *Genome*

Research **18,** 1043–1050 (2008).
22. Egert, M. *et al.* rRNA-based profiling of bacteria in the axilla of healthy males suggests right-left asymmetry in bacterial activity. *FEMS Microbiology Ecology* **77,** 146–153 (2011).
23. Kong, H. H. Skin microbiome: genomics-based insights into the diversity and role of skin microbes. *Trends in Molecular Medicine* **17,** 320–328 (2011).
24. Zeeuwen, P. L. J. M., Kleerebezem, M., Timmerman, H. M. & Schalkwijk, J. Microbiome and skin diseases. *Current Opinion in Allergy and Clinical Immunology* **13,** 514–520 (2013).
25. Fierer, N. *et al.* From the Cover: Forensic identification using skin bacterial communities. *PNAS* **107,** 6477–6481 (2010).
26. Dominguez-Bello, M. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *PNAS* **107,** 11971–11975 (2010).
27. Capone, K. A., Dowd, S. E., Stamatas, G. N. & Nikolovski, J. jid2011168a. *Journal of Investigative Dermatology* **131,** 2026–2032 (2011).
28. Marples, M. J. The ecology of the human skin. *The ecology of the human skin.* (1965).
29. leyden, J. J., McGinley, K. J., Hölzle, E., Labows, J. N. & Kligman, A. M. The microbiology of the human axilla and its relationship to axillary odor. *Journal of Investigative Dermatology* **77,** 413–416 (1981).
30. Lane, D. J. *et al.* Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences* **82,** 6955–6959 (1985).
31. Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *Journal of Bacteriology* **173,** 4371–4378 (1991).
32. Bond, P. L., Hugenholtz, P., Keller, J. & Blackall, L. L. Bacterial community structures of phosphate-removing and non-phosphate-removing activated sludges from sequencing batch reactors. *Applied and Environmental Microbiology* **61,** 1910–1916 (1995).
33. Ekendahl, S., Arlinger, J., Ståhl, F. & Pedersen, K. Characterisation of attached bacterial populations in deep granitic groundwater from the Stripa research mine by 16S rRNA gene sequencing and scanning electron microscopy. *Microbiology* **140,** 1575–1583 (1994).
34. Liesack, W. & Stackebrandt, E. Occurrence of novel groups of the domain Bacteria as revealed by analysis of genetic material isolated from an Australian terrestrial environment. *Journal of Bacteriology* **174,** 5072–5078 (1992).
35. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60-63 (1990).
36. DeLong, E. F., FRANKS, D. G. & ALLDREDGE, A. L. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Association for the Sciences of Limnology and Oceanography* **38** (1993).
37. Munson, M. A., Pitt-Ford, T., Chong, B., Weightman, A. & Wade, W. G. Molecular and Cultural Analysis of the Microflora Associated with Endodontic Infections. *Journal of Dental Research* **81,** 761–766 (2002).
38. Gao, Z., Tseng, C.-H., Pei, Z. & Blaser, M. J. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences* **104,** 2927–2932 (2007).
39. Edwards, R. A. *et al.* Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7,** 57 (2006).
40. Wegley, L., Edwards, R., Rodriguez-Brito, B., Liu, H. & Rohwer, F. Metagenomic analysis of the microbial community associated with the coral Porites astreoides. *Environ Microbiol* **9,** 2707–2719 (2007).
41. Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature*

452, 629–632 (2008).
42. Gomez-Alvarez, V., Teal, T. K. & Schmidt, T. M. Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal* **3,** 1314–1317 (2009).
43. Willner, D. *et al.* Metagenomic Analysis of Respiratory Tract DNA Viral Communities in Cystic Fibrosis and Non-Cystic Fibrosis Individuals. *PLoS ONE* **4,** e7370 (2009).
44. additional members, M. C. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
45. Lamendella, R., Santo Domingo, J. W., Ghosh, S., Martinson, J. & Oerther, D. B. Comparative fecal metagenomics unveils unique functional capacity of the swine gut. *BMC Microbiol* **11,** 103 (2011).
46. Belda-Ferre, P. *et al.* The oral metagenome in health and disease. *The ISME Journal* **6,** 46–56 (2011).
47. Yu, K. & Zhang, T. Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. *PLoS ONE* **7,** e38183 (2012).
48. Gill, S. R. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science* **312,** 1355–1359 (2006).
49. Yokouchi, H. *et al.* Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using phi29 polymerase. *Environ Microbiol* **8,** 1155–1163 (2006).
50. Costello, E. K. *et al.* Bacterial Community Variation in Human Body Habitats Across Space and Time. *Science* **326,** 1694–1697 (2009).
51. Fierer, N., Hamady, M., Lauber, C. L. & Knight, R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences* **105,** 17994–17999 (2008).
52. Grice, E. A. *et al.* Topographical and Temporal Diversity of the Human Skin Microbiome. *Science* **324,** 1190–1192 (2009).
53. Mathieu, A., Delmont, T. O., Vogel, T. M., Robe, P. & Nalin, R. Life on Human Surfaces: Skin Metagenomics. *PLoS ONE* (2013). doi:10.1371/journal.pone.0065288.g001
54. Meadow, J. F., Bateman, A. C., Herkert, K. M., O'Connor, T. K. & Green, J. L. Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ* **1,** e53 (2013).
55. Zeeuwen, P. L. *et al.* Microbiome dynamics of human epidermis following skin barrier disruption. *Genome Biology* **13,** R101 (2012).
56. Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2,** e00458–e00458 (2013).
57. Chiller, K., Selkin, B. A. & Murakawa, G. J. Skin microflora and bacterial infections of the skin. **6,** 170–174 (2001).
58. Cogen, A. L. *et al.* Selective Antimicrobial Action Is Provided by Phenol-Soluble Modulins Derived from Staphylococcus epidermidis, a Normal Resident of the Skin. *Journal of Investigative Dermatology* **130,** 192–200 (2010).
59. Iwase, T. *et al.* nature09074. *Nature* **465,** 346–349 (2010).
60. Wanke, I. *et al.* jid2010328a. *Journal of Investigative Dermatology* **131,** 382–390 (2010).
61. Lai, Y. *et al.* jid2010123a. *Journal of Investigative Dermatology* **130,** 2211–2221 (2010).
62. Li, D. *et al.* A Novel Lipopeptide from Skin Commensal Activates TLR2/CD36-p38 MAPK Signaling to Increase Antibacterial Defense against Bacterial Infection. *PLoS ONE* **8,** e58288 (2013).
63. Naik, S. *et al.* Compartmentalized Control of Skin Immunity by Resident Commensals. *Science* **337,** 1115–1119 (2012).

64. Lai, Y. *et al.* Commensal bacteria regulate Toll-like receptor 3–dependent inflammation after skin injury. *Nat Med* **15,** 1377–1382 (2009).
65. Bek-Thomsen, M., Lomholt, H. B. & Kilian, M. Acne is Not Associated with Yet-Uncultured Bacteria. *Journal of Clinical Microbiology* **46,** 3355–3360 (2008).
66. Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Research* **22,** 850–859 (2012).
67. Fahlén, A., Engstrand, L., Baker, B. S., Powles, A. & Fry, L. Comparison of bacterial microbiota in skin biopsies from normal and psoriatic skin. *Arch Dermatol Res* **304,** 15–22 (2011).
68. Gao, Z., Tseng, C.-H., Strober, B. E., Pei, Z. & Blaser, M. J. Substantial Alterations of the Cutaneous Bacterial Biota in Psoriatic Lesions. *PLoS ONE* **3,** e2719 (2008).
69. SHELLEY, W. B., HURLEY, H. J. J. & NICHOLS, A. C. Axillary odor: experimental study of the role of bacteria, apocrine sweat, and deodorants. *AMA Arch Derm Syphilol* **68,** 430–446 (1953).
70. SHEHADEH, N. H. & KLIGMAN, A. M. The effect of topical antibacterial agents on the bacterial flora of the axilla. *J Invest Dermatol* **40,** 61–71 (1963).
71. Tachibana, D. K. Microbiology of the foot. *Annual Reviews in Microbiology* **30,** 351–375 (1976).
72. Natsch, A., Derrer, S., Flachsmann, F. & Schmid, J. A Broad Diversity of Volatile Carboxylic Acids, Released by a Bacterial Aminoacylase from Axilla Secretions, as Candidate Molecules for the Determination of Human‐Body Odor Type. *Chemistry & biodiversity* **3,** 1–20 (2006).
73. James, A. G., Casey, J., Hyliands, D. & Mycock, G. Fatty acid metabolism by cutaneous bacteria and its role in axillary malodour. *World Journal of Microbiology and Biotechnology* **20,** 787–793 (2004).
74. Austin, C. & Ellis, J. Microbial pathways leading to steroidal malodour in the axilla. *The Journal of Steroid Biochemistry and Molecular Biology* **87,** 105–110 (2003).
75. Hasegawa, Y., Yabuki, M. & Matsukane, M. Identification of new odoriferous compounds in human axillary sweat. *Chemistry & biodiversity* **1,** 2042–2050 (2004).
76. Troccaz, M., Starkenmann, C., Niclass, Y., van de Waal, M. & Clark, A. J. 3‐Methyl‐3‐sulfanylhexan‐1‐ol as a Major Descriptor for the Human Axilla‐Sweat Odour Profile. *Chemistry & biodiversity* **1,** 1022–1035 (2004).
77. Ohloff, G., Maurer, B., Winter, B. & Giersch, W. Structural and configurational dependence of the sensory process in steroids. *Helvetica Chimica Acta* **66,** 192–217 (1983).
78. Decréau, R. A., Marson, C. M., Smith, K. E. & Behan, J. M. Production of malodorous steroids from androsta-5,16-dienes and androsta-4,16-dienes by Corynebacteria and other human axillary bacteria. *The Journal of Steroid Biochemistry and Molecular Biology* **87,** 327–336 (2003).
79. Schröder, J. *et al.* 1471-2164-13-141. *BMC Genomics* **13,** 141 (2012).
80. James, A. G., Hyliands, D. & Johnston, H. Generation of volatile fatty acids by axillary bacteria1. *International Journal of Cosmetic Science* **26,** 149–156 (2004).
81. Grice, E. A. & Segre, J. A. The skin microbiome. *Nat Rev Micro* **9,** 244–253 (2011).
82. Li, K., Bihan, M. & Methé, B. A. Analyses of the Stability and Core Taxonomic Memberships of the Human Microbiome. *PLoS ONE* **8,** e63139 (2013).
83. Götz, F., Verheij, H. M. & Rosenstein, R. Staphylococcal lipases: molecular characterisation, secretion, and processing. *Chemistry and physics of lipids* **93,**

15–25 (1998).
84. Rupp, M. E. & Archer, G. L. Coagulase-negative staphylococci: pathogens associated with medical progress. *CLIN INFECT DIS* 231–243 (1994).
85. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
86. Huang, J. T., Abrams, M., Tlougan, B., Rademaker, A. & Paller, A. S. Treatment of Staphylococcus aureus Colonisation in Atopic Dermatitis Decreases Disease Severity. *PEDIATRICS* **123,** e808–e814 (2009).
87. Staudinger, T., Pipal, A. & Redl, B. Molecular analysis of the prevalent microbiota of human male and female forehead skin compared to forearm skin and the influence of make-up. *Journal of Applied Microbiology* **110,** 1381–1389 (2011).
88. Natsch, A., Schmid, J. & Flachsmann, F. Identification of Odoriferous Sulfanylalkanols in Human Axilla Secretions and Their Formation through Cleavage of Cysteine Precursors by a CS Lyase Isolated from Axilla bacteria. *Chemistry & biodiversity* **1,** 1058–1072 (2004).
89. Marx, C. J. Getting in Touch with Your Friends. *Science* **324,** 1150–1151 (2009).
90. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host metabolism. *Nature* **489,** 242–249 (2012).
91. Iverson, V. *et al.* Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* **335,** 587–590 (2012).
92. Oh, S. *et al.* Metagenomic Insights into the Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Applied and Environmental Microbiology* **77,** 6000–6011 (2011).
93. Amann, R. I., Ludwig, W. & Schleifer, K.-H. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiology and Molecular Biology Reviews* **59,** 143–169 (1995).
94. Woese, C. R. Bacterial evolution. *Microbiological reviews* **51,** 221 (1987).
95. Woese, C. R., Stackebrandt, E., Macke, T. J. & Fox, G. E. A Phylogenetic Definition of the Major Eubacterial Taxa. *Systematic and Applied Microbiology* **6,** 143–151
96. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annual Reviews in Microbiology* **40,** 337–365 (1986).
97. Muyzer, G., De Waal, E. C. & Uitterlinden, A. G. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology* **59,** 695–700 (1993).
98. Ward, D. M., Weller, R. & Bateson, M. M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. (1990).
99. Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* **224,** 409–411 (1984).
100. Stahl, D. A., Lane, D. J., Olsen, G. J. & Pace, N. R. Characterisation of a Yellowstone hot spring microbial community by 5S rRNA sequences. *Applied and Environmental Microbiology* **49,** 1379–1384 (1985).
101. Lane, D. J., Stahl, D. A., Olsen, G. J., Heller, D. J. & Pace, N. R. Phylogenetic analysis of the genera Thiobacillus and Thiomicrospira by 5S rRNA sequences. *Journal of Bacteriology* **163,** 75–81 (1985).
102. McCabe, K. M., Zhang, Y.-H., Huang, B.-L., Wagar, E. A. & McCabe, E. R. Bacterial species identification after DNA amplification with a universal primer pair. *Molecular Genetics and Metabolism* **66,** 205–211 (1999).
103. Muyzer, G. & Waal, E. in *NATO ASI Series* (Stal, L. & Caumette, P.) **35,** 207–

214 (Springer Berlin Heidelberg, 1994).
104. Saiki, R. K. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239,** 487–491 (1988).
105. Kelly, J. J. *et al.* DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. *Water Research* **39,** 3229–3238 (2005).
106. DeSantis, T. Z. *et al.* High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microb Ecol* **53,** 371–383 (2007).
107. Massana, R., Murray, A. E., Preston, C. M. & DeLong, E. F. Vertical distribution and phylogenetic characterisation of marine planktonic Archaea in the Santa Barbara Channel. *Applied and Environmental Microbiology* **63,** 50–56 (1997).
108. Pace, N. R. A Molecular View of Microbial Diversity and the Biosphere. *Science* **276,** 734–740 (1997).
109. Zwart, G. *Divergent Members of the Bacterial Division Verrucomicrobiales in a Temperate Freshwater Lake*. (1998).
110. Dunbar, J., Takala, S., Barns, S. M., Davis, J. A. & Kuske, C. R. Levels of bacterial community diversity in four arid soils compared by cultivation and 16S rRNA gene cloning. *Applied and Environmental Microbiology* **65,** 1662–1669 (1999).
111. Orphan, V. J. *et al.* Comparative Analysis of Methane-Oxidizing Archaea and Sulfate-Reducing Bacteria in Anoxic Marine Sediments. *Applied and Environmental Microbiology* **67,** 1922–1934 (2001).
112. Oakley, B. B., Fiedler, T. L., Marrazzo, J. M. & Fredricks, D. N. Diversity of Human Vaginal Bacterial Communities and Associations with Clinically Defined Bacterial Vaginosis. *Applied and Environmental Microbiology* **74,** 4898–4909 (2008).
113. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences* **103,** 12115–12120 (2006).
114. DeSantis, T. Z. *et al.* Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology* **72,** 5069–5072 (2006).
115. Maidak, B. L. *et al*. The Ribosomal Database Project (RDP). *Nucleic Acids Research* **24**, 82-85 (1997).
116. Harrison, F. Microbial ecology of the cystic fibrosis lung. *Microbiology* **153,** 917–923 (2007).
117. Eckburg, P. B. Diversity of the Human Intestinal Microbial Flora. *Science* **308,** 1635–1638 (2005).
118. Jothibasu, K. Molecular Profiling of Rhizosphere Bacterial Communities Associated with Prosopis juliflora and Parthenium hysterophorus. *J. Microbiol. Biotechnol.* **22,** 301–310 (2012).
119. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2008).
120. Larsen, N. *et al.* Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults. *PLoS ONE* **5,** e9085 (2010).
121. Majumder, P. P. Genomic inferences on peopling of south Asia. *Current Opinion in Genetics & Development* **18,** 280–284 (2008).
122. Lagier, J. C. *et al.* Microbial culturomics: paradigm shift in the human gut microbiome study. *Clinical Microbiology and Infection* **18**, 1185-1193 (2012).
123. Wu, G. D. *et al.* Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334,** 105–108 (2011).
124. Dekio, I. Detection of potentially novel bacterial components of the human skin microbiota using culture-independent molecular profiling. *Journal of Medical Microbiology* **54,** 1231–1238 (2005).

125. Delmont, T. O. *et al.* Accessing the Soil Metagenome for Studies of Microbial Diversity. *Applied and Environmental Microbiology* **77,** 1315–1324 (2011).
126. Venter, J. C. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
127. Bond, P. L., Druschel, G. K. & Banfield, J. F. Comparison of acid mine drainage microbial communities in physically and geochemically distinct ecosystems. *Applied and Environmental Microbiology* **66,** 4962–4971 (2000).
128. Clement, B. G., Kehl, L. E., DeBord, K. L. & Kitts, C. L. Terminal restriction fragment patterns (TRFPs), a rapid, PCR-based method for the comparison of complex bacterial communities. *Journal of Microbiological Methods* **31,** 135–142 (1998).
129. Orita, M., Suzuki, Y., Sekiya, T. & Hayashi, K. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5,** 874–879 (1989).
130. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
131. Fraser Claire, M. *et al.* The minimal gene complement of Mycoplasma genitalium. (1995).
132. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18,** 821–829 (2008).
133. Lai, B., Ding, R., Li, Y., Duan, L. & Zhu, H. A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics* **28,** 1455–1462 (2012).
134. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98,** 9748–9753 (2001).
135. Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA–a practical iterative de Bruijn graph de novo assembler. 426–440 (2010).
136. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215,** 403–410 (1990).
137. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
138. Grim, C. J. *et al.* Pan-genome analysis of the emerging foodborne pathogen Cronobacter spp. suggests a species-level bidirectional divergence driven by niche adaptation. *BMC Genomics* **14,** 366 (2013).
139. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial '"pan-genome"'. *Proceedings of the Natural Academy of Sciences* **102,** 13950–13955 (2005).
140. Mann, R. A. *et al.* Comparative Genomics of 12 Strains of Erwinia amylovora Identifies a Pan-Genome with a Large Conserved Core. *PLoS ONE* **8,** e55644 (2013).
141. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Current Opinion in Genetics & Development* **15,** 589–594 (2005).
142. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology* **5,** R245–R249 (1998).
143. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
144. Strous, M. *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* **440,** 790–794 (2006).
145. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–131 (2006).
146. Rondon, M. R. *et al.* Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and Environmental Microbiology* **66,** 2541–2547 (2000).

147. Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. Characterisation of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* **178,** 591–599 (1996).
148. Beja, O. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* **289,** 1902–1906 (2000).
149. Abulencia, C. B. *et al.* Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments. *Applied and Environmental Microbiology* **72,** 3291–3301 (2006).
150. Lazarevic, V. *et al.* Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *Journal of Microbiological Methods* **79,** 266–271 (2009).
151. Cai, L. & Zhang, T. Detecting Human Bacterial Pathogens in Wastewater Treatment Plants by a High-Throughput Shotgun Sequencing Technique. *Environ. Sci. Technol.* **47,** 5433–5441 (2013).
152. Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences* **99,** 14250–14255 (2002).
153. Courtois, S. *et al.* Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ Microbiol* **3,** 431–439 (2001).
154. Ogram, A., Sayler, G. S. & Barkay, T. The extraction and purification of microbial DNA from sediments. *Journal of Microbiological Methods* **7,** 57–66 (1987).
155. Robe, P., Nalin, R., Capellano, C., Vogel, T. M. & Simonet, P. Extraction of DNA from soil. *European Journal of Soil Biology* **39,** 183–190 (2003).
156. Delmont, T. O., Robe, P., Clark, I., Simonet, P. & Vogel, T. M. Metagenomic comparison of direct and indirect soil DNA extraction approaches. *Journal of Microbiological Methods* **86,** 397–400 (2011).
157. Hunter, S. J. *et al.* Selective removal of human DNA from metagenomic DNA samples extracted from dental plaque. *J. Basic Microbiol.* **51,** 442–446 (2011).
158. Mullany, P., Hunter, S. & Allan, E. in *Advances in Applied Microbiology* **64,** 125–136 (Elsevier, 2008).
159. Foulongne, V. *et al.* Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing. *PLoS ONE* **7,** e38499 (2012).
160. Marine, R. *et al.* Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput Sequencing Libraries from Nanogram Quantities of DNA. *Applied and Environmental Microbiology* **77,** 8071–8079 (2011).
161. Duhaime, M. B., Deng, L., Poulos, B. T. & Sullivan, M. B. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* **14,** 2526–2537 (2012).
162. LaTuga, M. S. *et al.* Beyond Bacteria: A Study of the Enteric Microbial Consortium in Extremely Low Birth Weight Infants. *PLoS ONE* **6,** e27858 (2011).
163. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74,** 5463–5467 (1977).
164. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
165. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* (2005). doi:10.1038/nature03959
166. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242,** 84–89 (1996).

167. Adessi, C. *et al.* Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research* **28,** e87–e87 (2000).
168. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
169. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* **100,** 3960–3964 (2003).
170. Hoff, K. J. The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics* **10,** 520 (2009).
171. Nakamura, K. *et al.* Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research* **39,** e90–e90 (2011).
172. Niu, B., Fu, L., Sun, S. & Li, W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* **11,** 187 (2010).
173. Balzer, S., Malde, K., Grohme, M. A. & Jonassen, I. Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics* **29,** 830–836 (2013).
174. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Research* **9,** 868–877 (1999).
175. Myers, E. W. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology* **2,** 275–290 (1995).
176. Bonfield, J. K., Smith, K. F. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Research* **23,** 4992–4999 (1995).
177. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95,** 315–327 (2010).
178. Simpson, J. T. *et al.* ABySS: A parallel assembler for short read sequence data. *Genome Research* **19,** 1117–1123 (2009).
179. Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes. *Genome Research* **18,** 324–330 (2008).
180. MacCallum, I. *et al.* ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biology* **10,** R103 (2009).
181. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20,** 265–272 (2010).
182. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* **40,** e155–e155 (2012).
183. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).
184. Kultima, J. R. *et al.* MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS ONE* **7,** e47656 (2012).
185. Treangen, T. J. *et al.* gb-2013-14-1-r2. *Genome Biology* **14,** R2 (2013).
186. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology* **73,** 5261–5267 (2007).
187. Farrelly, V., Rainey, F. A. & Stackebrandt, E. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology* **61,** 2798–2801 (1995).
188. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods* **6,** 673–676 (2009).
189. Yang, B. *et al.* MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. 170–179 (2010).
190. McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I.

Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4,** 63–72 (2006).
191. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Research* **12,** 656–664 (2002).
192. Meyer, F. *et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
193. Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21,** 1552–1560 (2011).
194. Gerlach, W. & Stoye, J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research* **39,** e91–e91 (2011).
195. Krause, L. *et al.* Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* **36,** 2230–2239 (2008).
196. Gerlach, W., Jünemann, S., Tille, F., Goesmann, A. & Stoye, J. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* **10,** 430 (2009).
197. Lui, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic Profiling for Metagenomic Sequences. 1–6 (2012).
198. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9,** 811–814 (2012).
199. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Research* **38,** e191–e191 (2010).
200. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research* **38,** e132–e132 (2010).
201. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Research* **15,** 387–396 (2008).
202. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Research* **34,** 5623–5630 (2006).
203. Hoff, K. J., Lingner, T., Meinicke, P. & Tech, M. Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research* **37,** W101–W105 (2009).
204. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39,** W29–W37 (2011).
205. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research* **36,** D281–D288 (2007).
206. Hulo, N. The PROSITE database. *Nucleic Acids Research* **34,** D227–D230 (2006).
207. Attwood, T. K. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research* **31,** 400–402 (2003).
208. Yeats, C. *et al.* Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Research* **36,** D414–D418 (2007).
209. Haft, D. H. The TIGRFAMs database of protein families. *Nucleic Acids Research* **31,** 371–373 (2003).
210. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Research* **37,** D211–D215 (2009).
211. Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* **29,** 22–28 (2001).
212. Kaufmann, M. The role of the COG database in comparative and functional genomics. *Current Bioinformatics* **1,** 291–300 (2006).
213. Overbeek, R. The Subsystems Approach to Genome Annotation and its Use in

the Project to Annotate 1000 Genomes. *Nucleic Acids Research* **33,** 5691–5702 (2005).
214. Kanehisa, M. The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32,** 277D–280 (2004).
215. Kanehisa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* **34,** D354–D357 (2006).
216. Sun, S. *et al.* Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Research* **39,** D546–D551 (2010).
217. Lingner, T., Asshauer, K. P., Schreiber, F. & Meinicke, P. CoMet--a web server for comparative functional profiling of metagenomes. *Nucleic Acids Research* **39,** W518–W523 (2011).
218. Markowitz, V. M. *et al.* IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research* **40,** D123–D129 (2011).
219. Goll, J. *et al.* METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics. *Bioinformatics* **26,** 2631–2632 (2010).
220. Li, W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* **10,** 359 (2009).
221. Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J. & Bork, P. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* **26,** 2977–2978 (2010).
222. Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12,** 444 (2011).
223. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2,** 3 (2012).
224. Overbeek, R., Bartels, D., Vonstein, V. & Meyer, F. Annotation of Bacterial and Archaeal Genomes: Improving Accuracy and Consistency. *Chem. Rev.* **107,** 3431–3447 (2007).
225. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research* **40,** D130–D135 (2011).
226. Mavromatis, K. *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods* **4,** 495–500 (2007).
227. Mende, D. R. *et al.* Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. *PLoS ONE* **7,** e31386 (2012).
228. Jiang, H., An, L., Lin, S. M., Feng, G. & Qiu, Y. A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads. *PLoS ONE* **7,** e46450 (2012).
229. Tanaseichuk, O., Borneman, J. & Jiang, T. Separating metagenomic short reads into genomes via clustering. 298–313 (2011).
230. Morgan, J. L., Darling, A. E. & Eisen, J. A. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE* **5,** e10209 (2010).
231. Pignatelli, M. & Moya, A. Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. *PLoS ONE* **6,** e19984 (2011).
232. Morgan, X. C. & Huttenhower, C. Chapter 12: Human Microbiome Analysis. *PLoS Comput Biol* **8,** e1002808 (2012).
233. Telenius, H., Carter, N. P., Bebb, C. E., Ponder, B. A. & Tunnacliffe, A. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13,** 718–725 (1992).
234. Zhang, L. *et al.* Whole genome amplification from a single cell: Implications for geneticanalysis. *Proceedings of the Natural Academy of Sciences* **89,** 5847–5851 (1992).
235. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple

displacement amplification. *Proceedings of the National Academy of Sciences* **99,** 5261–5266 (2002).
236. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol* **7,** 19 (2007).
237. Kim, K. H. & Bae, J. W. Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Applied and Environmental Microbiology* **77,** 7663–7668 (2011).
238. Yilmaz, S., Allgaier, M. & Hugenholtz, P. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nature Methods* **7,** 943–944 (2010).
239. Syed, F., Grunenwald, H. & Caruccio, N. Optimised library preparation method for next-generation sequencing. *Nature Methods* **6,** i–ii (2009).
240. Parkinson, N. J. *et al.* Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Research* **22,** 125–133 (2012).
241. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology* **11,** R119 (2010).
242. Naumann, T. A. Tn5 Transposase Active Site Mutants. *Journal of Biological Chemistry* **277,** 17623–17629 (2002).
243. Edwards, U., Rogall, T., Blöcker, H., Emde, M. & Böttger, E. C. Isolation and direct complete nucleotide determination of entire genes. Characterisation of a gene coding for 16S ribosomal RNA. *Nucleic Acids Research* **17,** 7843–7853 (1989).
244. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25,** 0955–0964 (1997).
245. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8,** 209 (2007).
246. Markowitz, V. M. *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40,** D115–D122 (2011).
247. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research* **33,** W451–W454 (2005).
248. Haft, D. H. The TIGRFAMs database of protein families. *Nucleic Acids Research* **31,** 371–373 (2003).
249. Claudel-Renard, C. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research* **31,** 6633–6639 (2003).
250. Marchler-Bauer, A. *et al.* CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Research* **41,** D348–D352 (2012).
251. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
252. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
253. Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8,** 18 (2007).
254. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11,** 119 (2010).
255. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
256. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
257. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Research* **33,** W116–W120 (2005).

258. GO, T. Gene Ontology: tool for the unification of biology. *America* (2000).
259. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35,** 7188–7196 (2007).
260. The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Research* **40,** D71–D75 (2011).
261. Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research* **38,** D190–D195 (2009).
262. Sato, K., Kang, W. H., Saga, K. & Sato, K. T. Biology of sweat glands and their disorders. I. Normal sweat gland function. *Journal of American Dermatology* **20,** 537–563 (2013).
263. Labows, J. N., Preti, G., Hoelzle, E., leyden, J. J. & Kligman, A. M. STEROID ANALYSIS OF HUMAN APOCRINE SECRETION. *STEROIDS* **34,** 1–10 (1979).
264. Labows, J. N., McGinley, K. J. & Kligman, A. M. Perspectives on axillary odor. *Journal of the Society of Cosmetic Chemists* **34,** 193 (1982).
265. Patterson, M. J., Galloway, S. D. & Nimmo, M. A. Variations in regional sweat composition in normal human males. *Experimental Physiology* **85,** 869–875 (2000).
266. Bovell, D. L., Corbett, A. D., Holmes, S., MacDonald, A. & Harker, M. The absence of apoeccrine glands in the human axilla has disease pathogenetic implications, including axillary hyperhidrosis. *British Journal of Dermatology* **156,** 1278–1286 (2007).
267. Shehadeh, N. H. & Kligman, A. M. The Effect of Topical Antibacterial Agents on the Bacterial Flora of the Axilla1. *Journal of Investigative Dermatology* **40,** 61–71 (1963).
268. Taylor, D. *et al.* Characterisation of the microflora of the human axilla. *International Journal of Cosmetic Science* **25,** 137–145 (2003).
269. James, A. G., Austin, C. J., Cox, D. S., Taylor, D. & Calvert, R. Microbiological and biochemical origins of human axillary odour. *FEMS Microbiology Ecology* **83,** 527–540 (2013).
270. Gao, Z., Perez-Perez, G. I., Chen, Y. & Blaser, M. J. Quantitation of Major Human Cutaneous Bacterial and Fungal Populations. *Journal of Clinical Microbiology* **48,** 3575–3581 (2010).
271. Callewaert, C. *et al.* Characterisation of Staphylococcus and Corynebacterium Clusters in the Human Axillary Region. *PLoS ONE* **8,** e70538 (2013).
272. Rennie, P. J., Gower, D. B., holland, K. T., Mallet, A. I. & Watkins, W. J. The skin microflora and the formation of human axillary odour. *International Journal of Cosmetic Science* **12,** 197–207 (1990).
273. Gower, D. B., holland, K. T., Mallet, A. I., Rennie, P. J. & Watkins, W. J. Comparison of 16-Androstene steroid concentrations in sterile apocrine sweat and axillary secretions: Interconversions of 16-Androstenes by the axillary microflora—a mechanism for axillary odour production in man? *The Journal of Steroid Biochemistry and Molecular Biology* **48,** 409–418 (1994).
274. Amoore, J. E. Specific anosmia and the concept of primary odors. *Chemical Senses* **2,** 267–281 (1977).
275. Barzantny, H., Brune, I. & Tauch, A. Molecular basis of human body odour formation: insights deduced from corynebacterial genome sequences. *International Journal of Cosmetic Science* **34,** 2–11 (2011).
276. Petersen, L. J. Interstitial lactate levels in human skin at rest and during an oral glucose load: a microdialysis study. *Clinical Physiology* **19,** 246–250 (1999).
277. Thierry, A., Maillard, M. B. & Yvon, M. Conversion of L-Leucine to Isovaleric Acid by Propionibacterium freudenreichii TL 34 and ITGP23. *Applied and Environmental Microbiology* **68,** 608–615 (2002).

278. Natsch, A. A Specific Bacterial Aminoacylase Cleaves Odorant Precursors Secreted in the Human Axilla. *Journal of Biological Chemistry* **278,** 5718–5727 (2002).
279. Natsch, A., Gfeller, H., Gygax, P. & Schmid, J. Isolation of a bacterial enzyme releasing axillary malodor and its use as a screening target for novel deodorant formulations1. *International Journal of Cosmetic Science* **27,** 115–122 (2005).
280. Zeng, X.-N. *et al.* An investigation of human apocrine gland secretion for axillary odor precursors. *Journal of chemical ecology* **18,** 1039–1055 (2004).
281. zeng, C. *et al.* A human axillary odorant is carried by apolipoprotein D. *Proceedings of the National Academy of Sciences* **93,** 6626–6630 (1996).
282. Flower, D. The lipocalin protein family: structure and function. *Biochem. J* **318,** 1–14 (1996).
283. Troccaz, M. *et al.* Gender-Specific Differences between the Concentrations of Nonvolatile (R)/(S)-3-Methyl-3-Sulfanylhexan-1-Ol and (R)/(S)-3-Hydroxy-3-Methyl-Hexanoic Acid Odor Precursors in Axillary Secretions. *Chemical Senses* **34,** 203–210 (2009).
284. Akiba, S. *et al.* The N-terminal amino acid of apolipoprotein D is putatively covalently bound to 3-hydroxy-3-methyl hexanoic acid, a key odour compound in axillary sweat. *International Journal of Cosmetic Science* **33,** 283–286 (2011).
285. Preti, G. & leyden, J. J. Genetic Influences on Human Body Odor: From Genes to the Axillae. *Journal of Investigative Dermatology* **130,** 344–346 (2010).
286. Starkenmann, C., Niclass, Y., Troccaz, M. & Clark, A. J. Identification of the Precursor of (S)‐3‐Methyl‐3‐sulfanylhexan‐1‐ol, the Sulfury Malodour of Human Axilla Sweat. *Chemistry & biodiversity* **2,** 705–716 (2005).
287. Emter, R. & Natsch, A. The sequential action of a dipeptidase and a β-lyase is required for the release of the human body odorant 3-methyl-3-sulfanylhexan-1-ol from a secreted Cys-Gly-(S) conjugate by Corynebacteria. *Journal of Biological Chemistry* **283,** 20645–20652 (2008).
288. Martin, A. *et al.* jid2009254a. *Journal of Investigative Dermatology* **130,** 529–540 (2009).
289. Williamson, P. & Kligman, A. M. A new method for the quantitative investigation of cutaneous bacteria1. *Journal of Investigative Dermatology* **45,** 498–503 (1965).
290. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17,** pp. 10–12 (2011).
291. Rotmistrovsky, K. & Agarwala, R. BMTagger: Best Match Tagger for removing human reads from metagenomics datasets. (2011).
292. Schmieder, R. & Edwards, R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* **6,** e17288 (2011).
293. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
294. Caporaso, J. G. *et al.* QIIME allows analysis of high- throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).
295. Bray, J. R. & Curtis, J. T. An ordination of the upland forest communities of southern Wisconsin. *Ecological monographs* **27,** 325–349 (1957).
296. Parks, D. H. & Beiko, R. G. Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26,** 715–721 (2010).
297. Arndt, D. *et al.* METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Research* **40,** W88–W95 (2012).
298. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biology* **12,** R60 (2011).
299. Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* **47,** 583–621 (1952).

300. Wang, J. *et al.* Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease. *Sci. Rep.* **3,** (2013).
301. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
302. Cai, L., Ye, L., Tong, A. H. Y., Lok, S. & Zhang, T. Biased Diversity Metrics Revealed by Bacterial 16S Pyrotags Derived from Different Primer Sets. *PLoS ONE* **8,** e53649 (2013).
303. Ling, Z. *et al.* Pyrosequencing analysis of the human microbiota of healthy Chinese undergraduates. *BMC Genomics* **14,** 390 (2013).
304. Murdoch, D. A. Gram-positive anaerobic cocci. *Clinical Microbiology Reviews* **11,** 81–120 (1998).
305. Barzantny, H. *et al.* The transcriptional regulatory network of Corynebacterium jeikeium K411 and its interaction with metabolic routes contributing to human body odor formation. *Journal of Biotechnology* **159,** 235–248 (2012).
306. Tauch, A. *et al.* Ultrafast pyrosequencing of Corynebacterium kroppenstedtii DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids. *Journal of Biotechnology* **136,** 22–30 (2008).
307. Wauters, G., Van Bosterhaut, B., Janssens, M. & Verhaegen, J. Identification of Corynebacterium amycolatum and other nonlipophilic fermentative corynebacteria of human origin. *Journal of Clinical Microbiology* **36,** 1430–1432 (1998).
308. Graevenitz, A. & Bernard, K. in 819–842 (Springer New York, 2006). doi:10.1007/0-387-30743-5_31
309. Barreau, C., Bimet, F., Kiredjian, M., Rouillon, N. & Bizet, C. Comparative chemotaxonomic studies of mycolic acid-free coryneform bacteria of human origin. *Journal of Clinical Microbiology* **31,** 2085–2090 (1993).
310. Collins, M. D., Falsen, E., ÅKERVALL, E., SJÖDEN, B. & Alvarez, A. Note: Corynebacterium kroppenstedtii sp. nov., a novel corynebacterium that does not contain mycolic acids. *International journal of systematic bacteriology* **48,** 1449–1454 (1998).
311. Hall, V. Corynebacterium atypicum sp. nov., from a human clinical source, does not contain corynomycolic acids. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY* **53,** 1065–1068 (2003).
312. Takarada, H. *et al.* Complete Genome Sequence of the Soil Actinomycete Kocuria rhizophila. *Journal of Bacteriology* **190,** 4139–4146 (2008).
313. Kovács, G. *et al.* Kocuria palustris sp. nov. and Kocuria rhizophila sp. nov., isolated from the rhizoplane of the narrow-leaved cattail (Typha angustifolia). *International journal of systematic bacteriology* **49,** 167–173 (1999).
314. Becker, K. *et al.* Kocuria rhizophila Adds to the Emerging Spectrum of Micrococcal Species Involved in Human Infections. *Journal of Clinical Microbiology* **46,** 3537–3539 (2008).
315. Andersson, A. F. *et al.* Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. *PLoS ONE* **3,** e2836 (2008).
316. Binstock, J. F., Pramanik, A. & Schulz, H. Isolation of a multi-enzyme complex of fatty acid oxidation from Escherichia coli. *Proceedings of the National Academy of Sciences* **74,** 492–495 (1977).
317. Van der Meulen, R., Adriany, T., Verbrugghe, K. & De Vuyst, L. Kinetic Analysis of Bifidobacterial Metabolism Reveals a Minor Role for Succinic Acid in the Regeneration of NAD+ through Its Growth-Associated Production. *Applied and Environmental Microbiology* **72,** 5204–5210 (2006).
318. Teufel, L., Pipal, A., Schuster, K. C., Staudinger, T. & Redl, B. Material-dependent growth of human skin bacteria on textiles investigated using challenge tests and DNA genotyping. *Journal of Applied Microbiology* **108,**

450–461 (2010).
319. Blaser, M. J. *et al.* ismej201281a. **7,** 85–95 (2012).
320. Ezaki, T. *et al.* Proposal of the genera Anaerococcus gen. nov., Peptoniphilus gen. nov. and Gallicola gen. nov. for members of the genus Peptostreptococcus. *International journal of systematic bacteriology* **51,** 1521–1528 (2001).
321. Ezaki, T., Li, N. & Kawamura, Y. in 795–808 (Springer US, 2006). doi:10.1007/0-387-30744-3_26
322. Johnson, D. C., Dean, D. R., Smith, A. D. & Johnson, M. K. STRUCTURE, FUNCTION, AND FORMATION OF BIOLOGICAL IRON-SULFUR CLUSTERS. *Annu. Rev. Biochem.* **74,** 247–281 (2005).
323. Outten, F. W. The SufE Protein and the SufBCD Complex Enhance SufS Cysteine Desulfurase Activity as Part of a Sulfur Transfer Pathway for Fe-S Cluster Assembly in Escherichia coli. *Journal of Biological Chemistry* **278,** 45713–45719 (2003).
324. Schwartz, C. J., Djaman, O., Imlay, J. A. & Kiley, P. J. The cysteine desulfurase, IscS, has a major role in in vivo Fe-S cluster formation in Escherichia coli. *Proceedings of the National Academy of Sciences* **97,** 9009–9014 (2000).
325. Ayala-Castro, C., Saini, A. & Outten, F. W. Fe-S Cluster Assembly Pathways in Bacteria. *Microbiology and Molecular Biology Reviews* **72,** 110–125 (2008).
326. Agar, J. N. *et al.* IscU as a Scaffold for Iron−Sulfur Cluster Biosynthesis: Sequential Assembly of [2Fe-2S] and [4Fe-4S] Clusters in IscU †. *Biochemistry* **39,** 7856–7862 (2000).
327. Marienhagen, J., Kennerknecht, N., Sahm, H. & Eggeling, L. Functional Analysis of All Aminotransferase Proteins Inferred from the Genome Sequence of Corynebacterium glutamicum. *Journal of Bacteriology* **187,** 7639–7646 (2005).
328. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE* **6,** e27310 (2011).
329. Morales, S. E. & Holben, W. E. Empirical Testing of 16S rRNA Gene PCR Primer Pairs Reveals Variance in Target Specificity and Efficacy Not Suggested by In Silico Analysis. *Applied and Environmental Microbiology* **75,** 2677–2683 (2009).
330. Pester, M., Bittner, N., Deevong, P., Wagner, M. & Loy, A. ismej201075a. *The ISME Journal* 1–12 (2010). doi:10.1038/ismej.2010.75
331. Urich, T. *et al.* Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome. *PLoS ONE* **3,** e2527 (2008).
332. Poretsky, R. S. *et al.* Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11,** 1358–1375 (2009).
333. McCarren, J. *et al.* Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proceedings of the National Academy of Sciences* **107,** 16420–16427 (2010).
334. Hewson, I., Poretsky, R. S., Tripp, H. J., Montoya, J. P. & Zehr, J. P. Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ Microbiol* **12,** 1940–1956 (2010).
335. Otto, M. Staphylococcus epidermidis — the 'accidental' pathogen. *Nat Rev Micro* **7,** 555–567 (2009).
336. Cogen, A. L., Nizet, V. & Gallo, R. L. Skin microbiota: a source of disease or defence? *British Journal of Dermatology* **158,** 442–455 (2008).
337. Oh, J., Conlan, S., Polley, E. C., Segre, J. A. & Kong, H. H. Shifts in human

skin and nares microbiota of healthy children and adults. *Genome Medicine* **4,** 77 (2012).
338. Caputo, G. M., Archer, G. L., Calderwood, S. B., Dinubile, M. J. & Karchmer, A. W. Native valve endocarditis due to coagulase-negative staphylococci: clinical and microbiologic features. *The American journal of medicine* **83,** 619–625 (1987).
339. Foster, T. J. Immune evasion by staphylococci. *Nat Rev Micro* **3,** 948–958 (2005).
340. Iwatsuki, K., Yamasaki, O., Morizane, S. & Oono, T. Staphylococcal cutaneous infections: Invasion, evasion and aggression. *Journal of Dermatological Science* **42,** 203–214 (2006).
341. Gordon, R. J. & Lowy, F. D. Pathogenesis of Methicillin‐Resistant Staphylococcus aureusInfection. *CLIN INFECT DIS* **46,** S350–S359 (2008).
342. A report from the NNIS System. National Nosocomial Infections Surveillance (NNIS) System Report, data summary from January 1992 through June 2004, issued October 2004. *American Journal of Infection Control* **32,** 470–485 (2004).
343. Zhang, Y.-Q. *et al.* Genome-based analysis of virulence genes in a non-biofilm-forming Staphylococcus epidermidis strain (ATCC 12228). *Molecular Microbiology* **49,** 1577–1593 (2003).
344. Pitlik S, F. V. CEllulitis caused by staphylococcus epidermidis in a patient with leukemia. *Archives of Dermatology* **120,** 1099–1100 (1984).
345. Wang, A. *et al.* Contemporary clinical profile and outcome of prosthetic valve endocarditis. *JAMA: the journal of the American Medical Association* **297,** 1354–1361 (2007).
346. Diekema, D. J. *et al.* Survey of infections due to Staphylococcus species: frequency of occurrence and antimicrobial susceptibility of isolates collected in the United States, Canada, Latin America, Europe, and the Western Pacific region for the SENTRY Antimicrobial Surveillance Program, 1997–1999. *CLIN INFECT DIS* **32,** S114–S132 (2001).
347. Chambers, H. F., Hartman, B. J. & Tomasz, A. Increased amounts of a novel penicillin-binding protein in a strain of methicillin-resistant Staphylococcus aureus exposed to nafcillin. *J. Clin. Invest.* **76,** 325–331 (1985).
348. Knobloch, J. K. M. *et al.* Biofilm Formation by Staphylococcus epidermidis Depends on Functional RsbU, an Activator of the sigB Operon: Differential Activation Mechanisms Due to Ethanol and Salt Stress. *Journal of Bacteriology* **183,** 2624–2633 (2001).
349. Otto, M. Staphylococcal Infections: Mechanisms of Biofilm Maturation and Detachment as Critical Determinants of Pathogenicity*. *Annu. Rev. Med.* **64,** 175–188 (2013).
350. Mah, T.-F. C. & O'Toole, G. A. Mechanisms of biofilm resistance to antimicrobial agents. *Trends in Microbiology* **9,** 34–39 (2001).
351. Kristian, S. A. *et al.* Biofilm Formation Induces C3a Release and Protects Staphylococcus epidermidisfrom IgG and Complement Deposition and from Neutrophil‐Dependent Killing. *J INFECT DIS* **197,** 1028–1035 (2008).
352. Rohde, H. *et al.* Induction of Staphylococcus epidermidis biofilm formation via proteolytic processing of the accumulation-associated protein by staphylococcal and host proteases. *Molecular Microbiology* **55,** 1883–1895 (2005).
353. Heilmann, C. *et al.* Molecular basis of intercellular adhesion in the biofilm‐forming Staphylococcus epidermidis. *Molecular Microbiology* **20,** 1083–1091 (1996).
354. Mack, D., Haeder, M., Siemssen, N. & Laufs, R. Association of biofilm production of coagulase-negative staphylococci with expression of a specific polysaccharide intercellular adhesin. *Journal of Infectious Diseases* **174,** 881–

883 (1996).
355. Nilsson, M. *et al.* A fibrinogen-binding protein ofStaphylococcus epidermidis. *Infection and Immunity* **66,** 2666–2673 (1998).
356. Lou, Q. *et al.* Role of the SaeRS two-component regulatory system in Staphylococcus epidermidis autolysis and biofilm formation. *BMC Microbiol* **11,** 146 (2011).
357. Madhusoodanan, J. *et al.* An Enterotoxin-Bearing Pathogenicity Island in Staphylococcus epidermidis. *Journal of Bacteriology* **193,** 1854–1862 (2011).
358. Oliveira Calsolari, R. A. de, Pereira, V. C., Araújo Júnior, J. P. & de Souza da Cunha, M. de L. R. Determination of toxigenic capacity by reverse transcription polymerase chain reaction in coagulase-negative staphylococci and Staphylococcus aureus isolated from newborns in Brazil. *Microbiology and Immunology* **55,** 394–407 (2011).
359. Gill, S. R. *et al.* Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant Staphylococcus aureus Strain and a Biofilm-Producing Methicillin-Resistant Staphylococcus epidermidis Strain. *Journal of Bacteriology* **187,** 2426–2438 (2005).
360. The NIH HMP Working Group *et al.* The NIH Human Microbiome Project. *Genome Research* **19,** 2317–2323 (2009).
361. Conlan, S. *et al.* Staphylococcus epidermidis pan-genome sequence analysis reveals diversity of skin commensal and hospital infection-associated isolates. *Genome Biology* **13,** R64 (2012).
362. Nimmo, G. R. USA300 abroad: global spread of a virulent strain of community-associated methicillin-resistant Staphylococcus aureus. *Clinical Microbiology and Infection* **18,** 725–734 (2012).
363. Rammelkamp, C. H. & Maxon, T. Resistance of Staphylococcus aureus to the Action of Penicillin. **51,** 386–389 (1942).
364. Chambers, H. F. Methicillin resistance in staphylococci: molecular and biochemical basis and clinical implications. *Clinical Microbiology Reviews* **10,** 781–791 (1997).
365. Otto, M. MRSA virulence and spread. *Cell Microbiol* **14,** 1513–1521 (2012).
366. Mediavilla, J. R., Chen, L., Mathema, B. & Kreiswirth, B. N. Global epidemiology of community-associated methicillin resistant Staphylococcus aureus (CA-MRSA). *Current Opinion in Microbiology* **15,** 588–595 (2012).
367. Gal-Mor, O. & Finlay, B. B. Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol* **8,** 1707–1719 (2006).
368. Hansen, E. E. *et al.* Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences* **108,** 4599–4606 (2011).
369. Soares, S. C. *et al.* The Pan-Genome of the Animal Pathogen Corynebacterium pseudotuberculosis Reveals Differences in Genome Plasticity between the Biovar ovis and equi Strains. *PLoS ONE* **8,** e53818 (2013).
370. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* **11,** 472–477 (2008).
371. Tomida, S. *et al.* Pan-Genome and Comparative Genome Analyses of Propionibacterium acnes Reveal Its Genomic Diversity in the Healthy and Diseased Human Skin Microbiome. *mBio* **4,** e00003–13–e00003–13 (2013).
372. Ussery, D. W., Wassenaar, T. M. & Borini, S. in *Computational Biology* **8,** 213–228 (Springer London, 2009).
373. Read, T. D. & Ussery, D. W. Opening the pan-genomics box. *Current Opinion in Microbiology* **9,** 496–498 (2006).
374. Langille, M. G. I., Hsiao, W. W. L. & Brinkman, F. S. L. Detecting genomic islands using bioinformatics approaches. 1–10 (2010). doi:10.1038/nrmicro2350
375. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 Sequenced Escherichia coli Genomes. *Microb Ecol* **60,** 708–720 (2010).

376. Mavrodi, D. V., Loper, J. E., Paulsen, I. T. & Thomashow, L. S. Mobile genetic elements in the genome of the beneficial rhizobacterium Pseudomonas fluorescens Pf-5. *BMC Microbiol* **9,** 8 (2009).
377. Whittle, G., Shoemaker, N. B. & Salyers, A. A. The role of Bacteroides conjugative transposons in the dissemination of antibiotic resistance genes. *Cellular and Molecular Life Sciences CMLS* **59,** 2044–2054 (2002).
378. Hiller, N. L. *et al.* Comparative Genomic Analyses of Seventeen Streptococcus pneumoniae Strains: Insights into the Pneumococcal Supragenome. *Journal of Bacteriology* **189,** 8186–8195 (2007).
379. Rasko, D. A. *et al.* The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. *Journal of Bacteriology* **190,** 6881–6893 (2008).
380. Mira, A., Martín-Cuadrado, A. B., D'Auria, G. & Rodriguez-Valera, F. The bacterial pan-genome: a new paradigm in microbiology. *International Microbiology* **13,** 45–57 (2010).
381. Koonin, E. V., Makarova, K. S. & Aravind, L. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Annual Reviews in Microbiology* **55,** 709–742 (2001).
382. Dutta, C. & Pan, A. Horizontal gene transfer and bacterial diversity. *J Biosci* **27,** 27–33 (2002).
383. Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat Rev Micro* **3,** 722–732 (2005).
384. Salyers, A. A., Shoemaker, N. B., Stevens, A. M. & Li, L.-Y. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiology and Molecular Biology Reviews* **59,** 579–590 (1995).
385. Brüssow, H., Canchaya, C. & Hardt, W.-D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68,** 560–602 (2004).
386. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nat Rev Micro* **2,** 241–249 (2004).
387. Hacker, J., Hacker, J. H. & Kaper, J. B. *Pathogenicity Islands and the Evolution of Pathogenic Microbes*. (Springer Verlag, 2002).
388. Hsiao, W. W. L. *et al.* Evidence of a Large Novel Gene Pool Associated with Prokaryotic Genomic Islands. *PLoS Genet* **1,** e62 (2005).
389. Gans, J. Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. *Science* **309,** 1387–1390 (2005).
390. Curtis, T. P. MICROBIOLOGY: Exploring Microbial Diversity--A Vast Below. *Science* **309,** 1331–1333 (2005).
391. Goering, R. V. Pulsed field gel electrophoresis: A review of application and interpretation in the molecular epidemiology of infectious disease. *'Infection, Genetics and Evolution'* **10,** 866–875 (2010).
392. Davis, M. A., Hancock, D. D., Besser, T. E. & Call, D. R. Evaluation of Pulsed-Field Gel Electrophoresis as a Tool for Determining the Degree of Genetic Relatedness between Strains of Escherichia coli O157:H7. *Journal of Clinical Microbiology* **41,** 1843–1849 (2003).
393. Miragaia, M. *et al.* Comparison of Molecular Typing Methods for Characterisation of Staphylococcus epidermidis: Proposal for Clone Definition. *Journal of Clinical Microbiology* **46,** 118–129 (2008).
394. Janssen, P. *et al.* Evaluation of the DNA fingerprinting method AFLP as a new tool in bacterial taxonomy. *Microbiology* **142,** 1881–1893 (1996).
395. Vos, P. *et al.* AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23,** 4407–4414 (1995).
396. Zhao, S. *et al.* Genomic typing of Escherichia coli O157: H7 by semi-automated fluorescent AFLP analysis. *Microbes and Infection* **2,** 107–113 (2000).

397. Savelkoul, P. H. M. *et al.* High density whole genome fingerprinting of methicillin-resistant and -susceptible strains of Staphylococcus aureus in search of phenotype-specific molecular determinants. *Journal of Microbiological Methods* **71,** 44–54 (2007).
398. Stubbs, S. L., Brazier, J. S., O'Neill, G. L. & Duerden, B. I. PCR Targeted to the 16S-23S rRNA Gene Intergenic Spacer Region ofClostridium difficile and Construction of a Library Consisting of 116 Different PCR Ribotypes. *Journal of Clinical Microbiology* **37,** 461–463 (1999).
399. Killgore, G. *et al.* Comparison of seven techniques for typing international epidemic strains of Clostridium difficile: restriction endonuclease analysis, pulsed-field gel electrophoresis, PCR-ribotyping, multilocus sequence typing, multilocus variable-number tandem-repeat analysis, amplified fragment length polymorphism, and surface layer protein A gene sequence typing. *Journal of Clinical Microbiology* **46,** 431–437 (2008).
400. Wei, H. L., Kao, C. W., Wei, S. H., Tzen, J. T. & Chiou, C. S. Comparison of PCR ribotyping and multilocus variable-number tandem-repeat analysis (MLVA) for improved detection of Clostridium difficile. *BMC Microbiol* **11,** 217 (2011).
401. Kumar, N. S. & Gurusubramanian, G. Random amplified polymorphic DNA (RAPD) markers and its applications. (2011).
402. Lanini, S. *et al.* Molecular Epidemiology of a Pseudomonas aeruginosa Hospital Outbreak Driven by a Contaminated Disinfectant-Soap Dispenser. *PLoS ONE* **6,** e17064 (2011).
403. Chang, H. L. *et al.* Nosocomial Outbreak of Infection With Multidrug‐Resistant Acinetobacter baumanniiin a Medical Center in Taiwan •. *Infect Control Hosp Epidemiol* **30,** 34–38 (2009).
404. Maiden, M. C. *et al.* Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95,** 3140–3145 (1998).
405. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11,** 595 (2010).
406. Thomas, J. C. *et al.* Improved Multilocus Sequence Typing Scheme for Staphylococcus epidermidis. *Journal of Clinical Microbiology* **45,** 616–619 (2007).
407. Wang, X. M. Evaluation of a multilocus sequence typing system for Staphylococcus epidermidis. *Journal of Medical Microbiology* **52,** 989–998 (2003).
408. Mazars, E. *et al.* High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. *Proceedings of the National Academy of Sciences* **98,** 1901–1906 (2001).
409. Keim, P. *et al.* Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus anthracis. *Journal of Bacteriology* **182,** 2928–2936 (2000).
410. Titze-de-Almeida, R. *et al.* Multilocus Variable-Number Tandem-Repeat Polymorphism among Brazilian Enterococcus faecalis Strains. *Journal of Clinical Microbiology* **42,** 4879–4881 (2004).
411. Johansson, A., Koskiniemi, S., Gottfridsson, P., Wistrom, J. & Monsen, T. Multiple-Locus Variable-Number Tandem Repeat Analysis for Typing of Staphylococcus epidermidis. *Journal of Clinical Microbiology* **44,** 260–265 (2006).
412. Grissa, I., Bouchon, P., Pourcel, C. & Vergnaud, G. On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie* **90,** 660–668 (2008).
413. Denœud, F. & Vergnaud, G. Identification of polymorphic tandem repeats by

direct comparison of genome sequence from different bacterial strains : a web-based resource. *BMC Bioinformatics* **5,** 4 (2004).
414. Sabat, A., Malachowa, N., Miedzobrodzki, J. & Hryniewicz, W. Comparison of PCR-Based Methods for Typing Staphylococcus aureus Isolates. *Journal of Clinical Microbiology* **44,** 3804–3807 (2006).
415. Lindstedt, B.-A., Heir, E., Gjernes, E., Vardund, T. & Kapperud, G. DNA fingerprinting of Shiga-toxin producing *Escherichia coli* O157 based on Multiple-Locus Variable-Number Tandem-Repeats Analysis (MLVA). *Ann Clin Microbiol Antimicrob* **2,** 12 (2003).
416. Le Flèche, P. *et al.* A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. *BMC Microbiol* **1,** 2 (2001).
417. Pourcel, C., Vidgop, Y., Ramisse, F., Vergnaud, G. & Tram, C. Characterisation of a Tandem Repeat Polymorphism in Legionella pneumophila and Its Use for Genotyping. *Journal of Clinical Microbiology* **41,** 1819–1826 (2003).
418. Skuce, R. A. *et al.* Discrimination of Mycobacterium tuberculosis complex bacteria using novel VNTR-PCR targets. *Microbiology* **148,** 519–528 (2002).
419. Liao, J.-C., Li, C.-C. & Chiou, C.-S. Use of a multilocus variable-number tandem repeat analysis method for molecular subtyping and phylogenetic analysis of *Neisseria meningitidis* isolates. *BMC Microbiol* **6,** 44 (2006).
420. Spare, M. K. *et al.* Genotypic and phenotypic properties of coagulase-negative staphylococci causing dialysis catheter-related sepsis. *Journal of Hospital Infection* **54,** 272–278 (2003).
421. Marsh, J. W. *et al.* Multilocus variable-number tandem-repeat analysis for investigation of Clostridium difficile transmission in hospitals. *Journal of Clinical Microbiology* **44,** 2558–2566 (2006).
422. Moser, S. A. *et al.* Multiple-locus variable-number tandem-repeat analysis of meticillin-resistant Staphylococcus aureus discriminates within USA pulsed-field gel electrophoresis types. *Journal of Hospital Infection* **71,** 333–339 (2009).
423. Schleifer, K.-H. & Krämer, E. Selective Medium for Isolating Staphylococci. *Zentralblatt für Bakteriologie: I. Abt. Originale C: Allgemeine, angewandte und ökologische Mikrobiologie* **1,** 270–280
424. Schindler, C. A. & Schuhardt, V. T. Lysostaphin: a new bacteriolytic agent for the Staphylococcus. *Proceedings of the National Academy of Sciences of the United States of America* **51,** 414 (1964).
425. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341 (1987).
426. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37,** D141–D145 (2009).
427. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7,** e30619 (2012).
428. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Research* **38,** D211–D222 (2009).
429. Li, L. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13,** 2178–2189 (2003).
430. Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. 1471-2164-12-444-1. *BMC Genomics* **12,** 444 (2011).
431. White, J. R., Nagarajan, N. & Pop, M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol* **5,** e1000352 (2009).
432. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25,** 1335–1337 (2009).
433. Saeed, A. I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34,** 374–378 (2003).

434. McCrea, K. W. *et al.* The serine-aspartate repeat (Sdr) protein family in Staphylococcus epidermidis. *Microbiology* **146,** 1535–1546 (2000).
435. Egghe, L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *J. Am. Soc. Inf. Sci.* **58,** 702–709 (2007).
436. Bottacini, F. *et al.* Comparative genomics of the genus Bifidobacterium. *Microbiology* **156,** 3243–3254 (2010).
437. Budroni, S. *et al.* Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proceedings of the National Academy of Sciences* **108,** 4494–4499 (2011).
438. Trost, E. *et al.* Pangenomic Study of Corynebacterium diphtheriae That Provides Insights into the Genomic Diversity of Pathogenic Isolates from Cases of Classical Diphtheria, Endocarditis, and Pneumonia. *Journal of Bacteriology* **194,** 3199–3215 (2012).
439. Milani, C. *et al.* Comparative Genomics of Bifidobacterium animalis subsp. lactis Reveals a Strict Monophyletic Bifidobacterial Taxon. *Applied and Environmental Microbiology* **79,** 4304–4315 (2013).
440. The Human Microbiome Jumpstart Reference Strains Consortium *et al.* A Catalog of Reference Genomes from the Human Microbiome. *Science* **328,** 994–999 (2010).
441. Ludwig, W. ARB: a software environment for sequence data. *Nucleic Acids Research* **32,** 1363–1371 (2004).
442. Kather, B., Stingl, K., van der Rest, M. E., Altendorf, K. & Molenaar, D. Another Unusual Type of Citric Acid Cycle Enzyme inHelicobacter pylori: the Malate: Quinone Oxidoreductase. *Journal of Bacteriology* **182,** 3204–3209 (2000).
443. Beenken, K. E. *et al.* Global Gene Expression in Staphylococcus aureus Biofilms. *Journal of Bacteriology* **186,** 4665–4684 (2004).
444. Xue, T., You, Y., Hong, D., Sun, H. & Sun, B. The Staphylococcus aureus KdpDE Two-Component System Couples Extracellular K+ Sensing and Agr Signaling to Infection Programming. *Infection and Immunity* **79,** 2154–2167 (2011).
445. Kuroda, M. *et al.* Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in Staphylococcus aureus. *Molecular Microbiology* **49,** 807–821 (2004).
446. Gassel, M., Mollenkamp, T., Puppe, W. & Altendorf, K. The KdpF Subunit Is Part of the K+-translocating Kdp Complex of Escherichia coli and Is Responsible for Stabilization of the Complex in Vitro. *Journal of Biological Chemistry* **274,** 37901–37907 (1999).
447. Gries, C. M., Bose, J. L., Nuxoll, A. S., Fey, P. D. & Bayles, K. W. The Ktr potassium transport system in Staphylococcus aureusand its role in cell physiology, antimicrobial resistance and pathogenesis. *Molecular Microbiology* n/a–n/a (2013). doi:10.1111/mmi.12312
448. Bullen, J. J., Rogers, H. J., Spalding, P. B. & Ward, C. G. Iron and infection: the heart of the matter. *FEMS Immunology & Medical Microbiology* **43,** 325–330 (2005).
449. Ratledge, C. & Dover, L. G. Iron metabolism in pathogenic bacteria. *Annual Reviews in Microbiology* **54,** 881–941 (2000).
450. Maresso, A. W. & Schneewind, O. Iron Acquisition and Transport in Staphylococcus aureus. *Biometals* **19,** 193–203 (2006).
451. Dale, S. E., Sebulsky, M. T. & Heinrichs, D. E. Involvement of SirABC in Iron-Siderophore Import in Staphylococcus aureus. *Journal of Bacteriology* **186,** 8356–8362 (2004).
452. Reniere, M. L. & Skaar, E. P. Staphylococcus aureushaem oxygenases are differentially regulated by iron and haem. *Molecular Microbiology* **69,** 1304–1315 (2008).

453. Skaar, E. P. IsdG and IsdI, Heme-degrading Enzymes in the Cytoplasm of Staphylococcus aureus. *Journal of Biological Chemistry* **279,** 436–443 (2003).
454. Mazmanian, S. K. Passage of Heme-Iron Across the Envelope of Staphylococcus aureus. *Science* **299,** 906–909 (2003).
455. Hurd, A. F. *et al.* The iron-regulated surface proteins IsdA,IsdB, and IsdH are not required for heme iron utilization in Staphylococcus aureus. *FEMS Microbiology Letters* **329,** 93–100 (2012).
456. Clarke, S. R. & Foster, S. J. in *Advances in Microbial Physiology* **51,** 187–224 (Elsevier, 2006).
457. Clarke, S. R. *et al.* The Staphylococcus aureus Surface Protein IsdA Mediates Resistance to Innate Defenses of Human Skin. *Cell Host and Microbe* **1,** 199–212 (2007).
458. Visai, L. *et al.* Immune evasion by Staphylococcus aureus conferred by iron-regulated surface determinant protein IsdH. *Microbiology* **155,** 667–679 (2009).
459. Andrews, S. C., Robinson, A. K. & RodrÃ guez-QuiÃ ones, F. Bacterial iron homeostasis. *FEMS Microbiology Reviews* **27,** 215–237
460. KONETSCHNY RAPP, S., Jung, G., MEIWES, J. & ZÄHNER, H. Staphyloferrin A: a structurally new siderophore from staphylococci. *European Journal of Biochemistry* **191,** 65–74 (1990).
461. Drechsel, H. *et al.* Purification and chemical characterisation of staphyloferrin B, a hydrophilic siderophore from staphylococci. *Biometals* **6,** 185–192 (1993).
462. Courcol, R. J., Trivier, D., Bissinger, M.-C., Martin, G. R. & Brown, M. R. Siderophore production by Staphylococcus aureus and identification of iron-regulated proteins. *Infection and Immunity* **65,** 1944–1948 (1997).
463. Dale, S. E., Doherty-Kirby, A., Lajoie, G. & Heinrichs, D. E. Role of Siderophore Biosynthesis in Virulence of Staphylococcus aureus: Identification and Characterisation of Genes Involved in Production of a Siderophore. *Infection and Immunity* **72,** 29–37 (2003).
464. Cheung, J., Beasley, F. C., Liu, S., Lajoie, G. A. & Heinrichs, D. E. Molecular characterisation of staphyloferrin B biosynthesis in Staphylococcus aureus. *Molecular Microbiology* **74,** 594–608 (2009).
465. Bhatt, G. & Denny, T. P. Ralstonia solanacearum Iron Scavenging by the Siderophore Staphyloferrin B Is Controlled by PhcA, the Global Virulence Regulator. *Journal of Bacteriology* **186,** 7896–7904 (2004).
466. Beasley, F. C., Cheung, J. & Heinrichs, D. E. 1471-2180-11-199. *BMC Microbiol* **11,** 199 (2011).
467. Kocianova, S. *et al.* Key role of poly-γ-dl-glutamic acid in immune evasion and virulence of Staphylococcus epidermidis. *J. Clin. Invest.* **115,** 688–694 (2005).
468. Thakker, M., Park, J. S., Carey, V. & Lee, J. C. Staphylococcus aureus serotype 5 capsular polysaccharide is antiphagocytic and enhances bacterial virulence in a murine bacteremia model. *Infection and Immunity* **66,** 5183–5189 (1998).
469. Arbeit, R. D., Karakawa, W. W., Vann, W. F. & Robbins, J. B. Predominance of two newly described capsular polysaccharide types among clinical isolates of Staphylococcus aureus. *Diagnostic microbiology and infectious disease* **2,** 85–91 (1984).
470. Verdier, I. *et al.* Identification of the Capsular Polysaccharides in Staphylococcus aureus Clinical Isolates by PCR and Agglutination Tests. *Journal of Clinical Microbiology* **45,** 725–729 (2007).
471. Sau, S. *et al.* The Staphylococcus aureus allelic genetic loci for serotype 5 and 8 capsule expression contain the type-specific genes flanked by common genes. *Microbiology* **143,** 2395–2405 (1997).
472. Jones, C. Revised structures for the capsular polysaccharides from Staphylococcus aureus Types 5 and 8, components of novel glycoconjugate vaccines. *Carbohydrate Research* **340,** 1097–1106 (2005).
473. Kiser, K. B., Bhasin, N., Deng, L. & Lee, J. C. Staphylococcus aureus cap5P

encodes a UDP-N-acetylglucosamine 2-epimerase with functional redundancy. *Journal of Bacteriology* **181,** 4818–4824 (1999).
474. Cramton, S. E., Gerke, C., Schnell, N. F., Nichols, W. W. & Götz, F. The intercellular adhesion (ica) locus is present in Staphylococcus aureus and is required for biofilm formation. *Infection and Immunity* **67,** 5427–5433 (1999).
475. Cafiso, V. *et al.* Presence of the ica operon in clinical isolates of Staphylococcus epidermidis and its role in biofilm production. *Clinical Microbiology and Infection* **10,** 1081–1088 (2004).
476. Deighton, M. A., Franklin, J. C., Spicer, W. J. & Balkau, B. Species identification, antibiotic sensitivity and slime production of coagulase-negative staphylococci isolated from clinical specimens. *Epidemiology and infection* **101,** 99–113 (1988).
477. Jefferson, K. K., Pier, D. B., Goldmann, D. A. & Pier, G. B. The Teicoplanin-Associated Locus Regulator (TcaR) and the Intercellular Adhesin Locus Regulator (IcaR) Are Transcriptional Inhibitors of the ica Locus in Staphylococcus aureus. *Journal of Bacteriology* **186,** 2449–2456 (2004).
478. Conlon, K. M., Humphreys, H. & O'Gara, J. P. icaR Encodes a Transcriptional Repressor Involved in Environmental Regulation of ica Operon Expression and Biofilm Formation in Staphylococcus epidermidis. *Journal of Bacteriology* **184,** 4400–4408 (2002).
479. Rachid, S. *et al.* Alternative Transcription Factor ςB Is Involved in Regulation of Biofilm Expression in a Staphylococcus aureusMucosal Isolate. *Journal of Bacteriology* **182,** 6824–6826 (2000).
480. Cramton, S. E., Ulrich, M., Gotz, F. & Doring, G. Anaerobic Conditions Induce Expression of Polysaccharide Intercellular Adhesin in Staphylococcus aureus and Staphylococcus epidermidis. *Infection and Immunity* **69,** 4079–4085 (2001).
481. Vuong, C., Saenz, H. L., Götz, F. & Otto, M. Impact of the agr quorum-sensing system on adherence to polystyrene in Staphylococcus aureus. *Journal of Infectious Diseases* **182,** 1688–1693 (2000).
482. Arciola, C. R., Baldassarri, L. & Montanaro, L. Presence of icaA and icaD Genes and Slime Production in a Collection of Staphylococcal Strains from Catheter-Associated Infections. *Journal of Clinical Microbiology* **39,** 2151–2156 (2001).
483. Chang, Y.-M. *et al.* Structural study of TcaR and its complexes with multiple antibiotics from Staphylococcus epidermidis. *Proceedings of the National Academy of Sciences* **107,** 8617–8622 (2010).
484. Kumar, P. S., Brooker, M. R., Dowd, S. E. & Camerlengo, T. Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. *PLoS ONE* **6,** e20956 (2011).
485. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21,** 494–504 (2011).
486. Diep, B. A. *et al.* Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *The Lancet* **367,** 731–739 (2006).
487. Hanssen, A. M., Kjeldsen, G. & Sollid, J. U. E. Local Variants of Staphylococcal Cassette Chromosome mec in Sporadic Methicillin-Resistant Staphylococcus aureus and Methicillin-Resistant Coagulase-Negative Staphylococci: Evidence of Horizontal Gene Transfer? *Antimicrobial Agents and Chemotherapy* **48,** 285–296 (2003).
488. Myllykangas, S., Buenrostro, J. and Hanlee, P. Ji. Bioinformatics for High Throughput Sequencing, Overview of Sequencing Technology Platforms. Springer New York (eds Rodríguez-Ezpeleta, N., Hackenberg, M., Aransay, A.M.) pp 11-25 (Springer New York, 2012).

489. Schmieder, R and Edwards, R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLoS ONE* **6** , e17288 (2011).